

AD-A273 497



AFOSR-TR- 93 0878

ANNUAL/FINAL TECHNICAL REPORT

PHOTONIC TECHNOLOGY FOR  
IMPLEMENTATION OF GENERALIZABLE  
NEURAL NETWORKS: A SYNTHETIC  
APPROACH

DARPA/AFOSR Grant No. AFOSR-89-0466

submitted by:

B. Keith Jenkins, Anupam Madhukar and Armand R. Tanguay, Jr.

2

University of Southern California

Signal and Image Processing Institute

Department of Electrical Engineering-Systems  
3740 McClintock Ave., Room 400  
Los Angeles, CA 90089-2564

S DTIC  
ELECTE  
DEC 07 1993  
A



This document has been approved  
for public release; distribution is unlimited.

93-29782



93 12 6 072

**Best  
Available  
Copy**

2

**ANNUAL/FINAL TECHNICAL REPORT**

**PHOTONIC TECHNOLOGY FOR  
IMPLEMENTATION OF GENERALIZABLE  
NEURAL NETWORKS: A SYNTHETIC  
APPROACH**

DARPA/AFOSR Grant No. AFOSR-89-0466

submitted by:

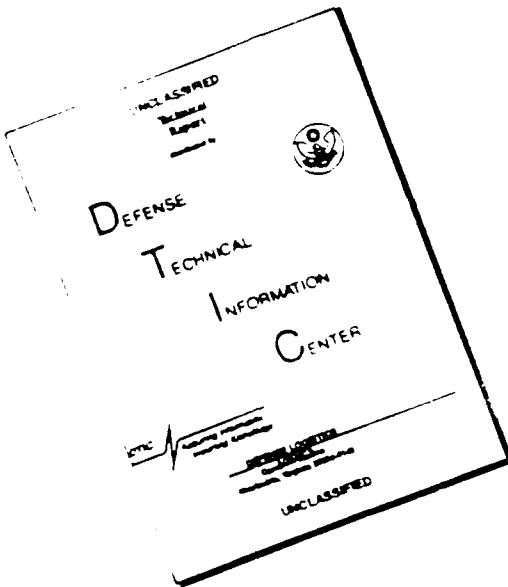
B. Keith Jenkins, Anupam Madhukar and Armand R. Tanguay, Jr.

DTIC  
SELECTED  
DEC 07 1993  
S A

This document has been approved  
for public release and sale; its  
distribution is unlimited.

Accession For	
NTIS	CRA&I
DTIC	TAB
Unannounced	
Justification	
By _____	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

# **DISCLAIMER NOTICE**



**THIS DOCUMENT IS BEST  
QUALITY AVAILABLE. THE COPY  
FURNISHED TO DTIC CONTAINED  
A SIGNIFICANT NUMBER OF  
PAGES WHICH DO NOT  
REPRODUCE LEGIBLY.**

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

The reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Security, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-3182, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

			Form Approved OMB No. 0704-0188
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		October 28, 1993	Final Technical Report (9/30/89 - 9/15/93)
4. TITLE AND SUBTITLE			5. FUNDING NUMBERS
Photonic Technology for Implementation of Generalizable Neural Networks: A Synthetic Approach			Grant No. AFOSR890466
6. AUTHOR(S)			
B. Keith Jenkins, P.I. Anupam Madhukar, co-P.I. Armand R. Tanguay, Jr., co-P.I.			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)			8. PERFORMING ORGANIZATION REPORT NUMBER
University of Southern California Department of Electrical Engineering/Electrophysics Signal and Image Processing Institute 3740 McClintock Avenue, Suite 404 Los Angeles, CA 90089-2564			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSORING/MONITORING AGENCY REPORT NUMBER
Alan E. Craig AFOSR/NE Bolling AFB, Bldg. 410 Washington, DC 20332-6448			Darpa
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION/AVAILABILITY STATEMENT			12b. DISTRIBUTION CODE
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED			
13. ABSTRACT (Maximum 200 words)			
<p>This report describes progress and results on the grant, "Photonic Technology for Implementation of Generalizable Neural Networks: A Synthetic Approach," DARPA/AFOSR Grant No. AFOSR-89-0466, for the annual report period 30 Sept 91 - 29 Sept 92, and for the final reporting period 30 Sept 89 - 15 Sept 93. Substantial progress and results have been achieved in the areas of neural network models for optoelectronic systems, photonic neural network architectures, incoherent/coherent holographic interconnections, optical disk spatial light modulators with parallel readout (as input devices), and high bandwidth spatial light modulators (for neuron bit arrays) including mutually compatible silicon detectors, silicon electronics, and high contrast ration InGaAs/GaAs modulators.</p>			
14. SUBJECT TERMS			15. NUMBER OF PAGES
			70
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT		18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED		INCLASSIFIED	UNCLASSIFIED
20. LIMITATION OF ABSTRACT			UL

**Photonic Technology for Implementation of Generalizable Neural Networks:  
A Synthetic Approach**

DARPA/AFOSR Grant No. AFOSR-89-0466

Combined

Annual Technical Report, Reporting Period: 30 Sept 91 - 29 Sept 92

Final Technical Report, Reporting Period: 30 Sept 89 - 15 Sept 93

B. Keith Jenkins, P.I.

Anupam Madhukar, co-P.I.

Armand R. Tanguay, Jr., co-P.I.

Departments of Electrical Engineering and Materials Science

Center for Photonic Technology,

Center for Neural Engineering, and Signal and Image Processing Institute

University of Southern California

Los Angeles, California 90089-2564

Submitted 27 October 1993

## **Table of Contents**

<b>1. Executive Summary</b>	<b>3</b>
<b>2. Summary of Approach, Progress, and Results</b>	<b>5</b>
<b>2.1 Introduction</b>	
<b>2.2 Photonic Architecture</b>	
<b>2.3 Photonic Components</b>	
<b>2.4 References</b>	
<b>3. List of Accomplishments</b>	<b>12</b>
<b>3.1 Multilayer Learning Algorithm</b>	
<b>3.2 Photonic Neural Network Architecture</b>	
<b>3.3 Incoherent/Coherent Fan-in/Fan-out Interconnection System</b>	
<b>3.4 Silicon Detector and Analog Circuitry Arrays</b>	
<b>3.5 GaAs/InGaAs Modulator Arrays</b>	
<b>3.6 Incoherent/Coherent Source Arrays</b>	
<b>3.7 Parallel Readout Optical Disk</b>	
<b>4. Publications and Patents to Date Resulting from this Grant</b>	<b>15</b>
<b>5. Copies of Selected Publications for Further Information</b>	<b>20</b>

## **1. Executive Summary**

This report describes progress and results on the grant, "Photonic Technology for Implementation of Generalizable Neural Networks: A Synthetic Approach", DARPA/AFOSR Grant No. AFOSR-89-0466, for the annual reporting period 30 Sept 91 - 29 Sept 92, and for the final reporting period 30 Sept 89 - 15 Sept 93. Substantial progress and results have been achieved in the areas of neural network models for optoelectronic systems, photonic neural network architectures, incoherent/coherent holographic interconnections, optical disk spatial light modulators with parallel readout (as input devices), and high bandwidth spatial light modulators (for neuron unit arrays) including mutually compatible silicon detectors, silicon electronics, and high contrast ratio InGaAs/GaAs modulators.

Our approach has been vertically integrated in that it synthesizes materials, devices, architectures, and algorithms. The effort has been centered around a novel class of photonic architectures that provide for generalizability to different neuron and neural network models, modularity, fully parallel input/output, and scalability to large numbers of neuron units and interconnections. The architectures incorporate optoelectronic spatial light modulators for flexible neuron unit functionality by means of integrated silicon electronics, as well as silicon detectors and GaAs-based modulator elements for highly parallel input to, and output from, each neuron unit array. Each pixel of this neuron unit array includes dual channel inputs and outputs to enable bipolar functionality. For consistency among the algorithm/model level, the architecture level, and the hardware component and materials levels, we have developed a multilayer neural network model and learning algorithm based on this dual channel signal representation. Furthermore, we have maintained compatibility of all hardware architectures, devices, and materials with this model.

One of the key results of this effort has been the simultaneous development of mutually compatible high-contrast GaAs-based modulator element arrays for neuron unit output, silicon detector and electronic circuitry arrays for neuron unit input and signal processing, and multilayer neural network models and learning algorithms. Both device arrays have been fabricated and demonstrated to yield high

contrast, high bandwidth, high density, and dual channel capability that is compatible with the dual channel neural network model mentioned above.

A two dimensional array of surface-emitting semiconductor laser diodes can provide a compact implementation of the requisite sources; for laboratory purposes we have utilized other novel, programmable techniques for generation of source arrays at visible as well as infrared wavelengths. In addition, since scalability places demands not only on the interconnection and the neuron unit arrays, but also on the input device, we have investigated the use of an optical disk spatial light modulator that provides parallel readout of a large number of stored input training patterns at a high frame rate.

Central to our photonic architecture is a novel volume holographic interconnection technique that is based on an array of individually coherent but mutually incoherent sources. Results from our experimental, computational, and analytical investigations of the properties of this interconnection system have been a key impetus in the pursuit of this overall work, and have shown the interconnection system to be exceptionally low in crosstalk, despite highly parallel recording and readout of the volume holographic element. This photonic interconnection architecture exhibits another unique property: once trained, the stored weighted-interconnection pattern can be copied in a single exposure step to another volume holographic medium, thus permitting duplication of a pretrained neural network in a manufacturing environment.

## **2. Summary of Approach, Progress, and Results**

This section gives an overview of our approach, progress, and results. Detailed descriptions are given in the attached papers and in the previous annual reports, including the papers attached therein. A list of accomplishments is given in Section 3.

### **2.1 Introduction**

A general neural network architecture should have the following features: (1) modularity, i.e., be in the form of a cascadable "module"; (2) capability for lateral, feedforward, and feedback interconnections; (3) analog weighted interconnections; (4) bipolar signals and weights; (5) scalability to large numbers of neuron units with high connectivity; (6) generalizability to different network models and learning algorithms, as well as capability for extension to possible future network models.

Recently, optical and optoelectronic implementations of neural networks have been the focus of intensive research and development (see for example, Ref. 1). In our effort we have developed a novel implementation technique that provides for most if not all of the above properties. Our photonic architecture<sup>2</sup> utilizes a novel incoherent/coherent hologram recording and reconstruction technique for highly parallel operation with reduced crosstalk. High bandwidth optoelectronic spatial light modulators with integrated detectors, modulators, and control electronics at each pixel are being fabricated for the 2-D neuron unit arrays.

### **2.2 Photonic Architecture**

A broad class of learning algorithms can be represented by the weight update equation:

$$\Delta w_{ij} = \alpha \delta_i x_j - \beta w_{ij}$$

in which  $w_{ij}$  is the interconnection weight from  $j$  to  $i$ ,  $\Delta w_{ij} = \Delta w_{ij}(k+1) - w_{ij}(k)$  is the weight update,  $\alpha$  is the learning gain constant,  $x_j$  is the signal level of the  $j^{\text{th}}$  input (e.g.,  $j^{\text{th}}$  neuron unit of the previous layer in a multilayer net), and  $\beta$  is the decay constant. Suitable choices of  $\delta_i$  give different learning algorithms, such as Hebbian, Widrow-Hoff, and back propagation.

In our photonic architecture the neural interconnections as well as the weight updates that train the interconnections are formed within a photorefractive crystal. The fanout interconnection pattern from one neuron unit is implemented as a single hologram,  $N^2$  of which are multiplexed into a single volume holographic medium. The last term in the above equation is an optional decay term that is included primarily to model intentional or unintentional decay of holograms in such a photorefractive crystal. (Other physical effects such as nonlinearities and saturation in the medium are considered entirely unintentional and are treated separately.)

Two-dimensional optoelectronic GaAs-based spatial light modulator (SLM) arrays, with integrated detectors, modulators and control electronics, implement inner product neuron units, as well as the training plane array, which generates the  $\delta_i$  terms. This technology provides for: (1) incorporation of bipolar signals via dual-channel inputs and outputs; (2) slight variants of the same basic SLM structure for all SLM's in the architecture (for neuron units and  $\delta_i$  generation); (3) incorporation of different neuron unit functions, including linear, soft threshold, and hard clipping, as well as variable gain; and (4) potential extendibility to future neural network models.

The photonic architecture for the case of Hebbian learning,  $\delta_i = y_i$ , where  $y_i$  is the output of neuron unit  $i$ , is shown in Fig. 1. (For other learning algorithms, SLM<sub>1</sub> is replaced by one or more SLM's with appropriate functions, and for supervised learning another beam representing desired outputs is incident on the SLM.) Only feedforward connections are shown.

The neuron unit arrays mentioned above are incorporated by means of two spatial light modulators (SLM's). SLM<sub>2</sub> represents the array of input units to the

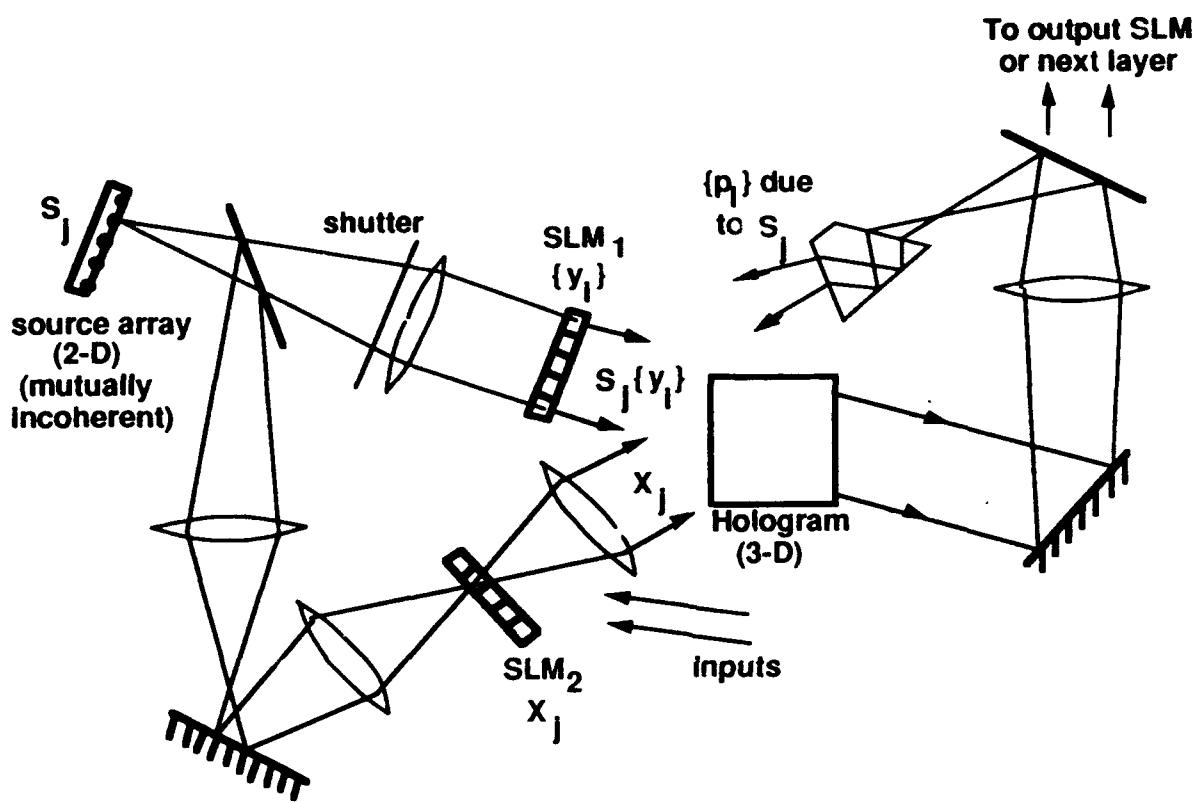


Figure 1. Photonic architecture for neural networks, Hebbian case.

current optical module, and  $SLM_1$  represents the array of output neuron units of the current module (yielding outputs of the neural net or providing the input neuron units of the next module). An array of mutually incoherent coherent sources is used to illuminate the system; these sources are provided by a mutually incoherent laser diode array, or by a coherent beam passing through an optical system that generates a set of beams, each temporally modulated or frequency shifted sufficiently to provide mutual incoherence over the detection time constant.

During the computation phase, the shutter is closed to prevent learning. The array of sources is imaged onto  $SLM_2$  as a set of read beams. Each beam is modulated by  $SLM_2$  to provide the inputs. Then each (reference) beam  $x_j$  passes through the hologram to provide the interconnections, i.e., the weighted fanout from each input  $x_j$ . The hologram output is sent to the write side of  $SLM_1$  (i.e., the virtual image of  $SLM_1$  from the reconstruction is imaged onto  $SLM_1$  via the lens). Optical fan-in of multiple holographic reconstructions at this plane provides linear incoherent summation of weighted interconnections at each neuron unit input. The dove prism reflects the images so that their orientation at  $SLM_1$  is consistent.

In the learning phase, the shutter is open. Each weight update term is computed optically and recorded into the photorefractive material. Light from each source is approximately collimated and used as the read beam for  $SLM_1$ . Thus, for an  $N$  by  $N$  array of sources, there are  $N^2$  beams reading  $SLM_1$  simultaneously, each at a different angle, as shown in Fig. 2; all  $y_i$  terms are encoded onto each of these beams. This novel angular multiplexing technique eliminates beam degeneracy, thereby reducing crosstalk during readout compared with more conventional single-source fully coherent recording techniques<sup>3</sup>. Each of these beams then interferes only with its corresponding reference beam,  $x_j$ , from the same source, in the photorefractive material. This process writes the set of desired weight update terms  $\alpha x_j y_i$ .

The central architectural feature of *incoherent/coherent double angularly multiplexed* hologram recording and readout has been designed to satisfy a number of constraints: minimization of inherent recording-induced crosstalk; minimization of inherent readout-induced (fan-in) crosstalk; incoherent (additive) summation of

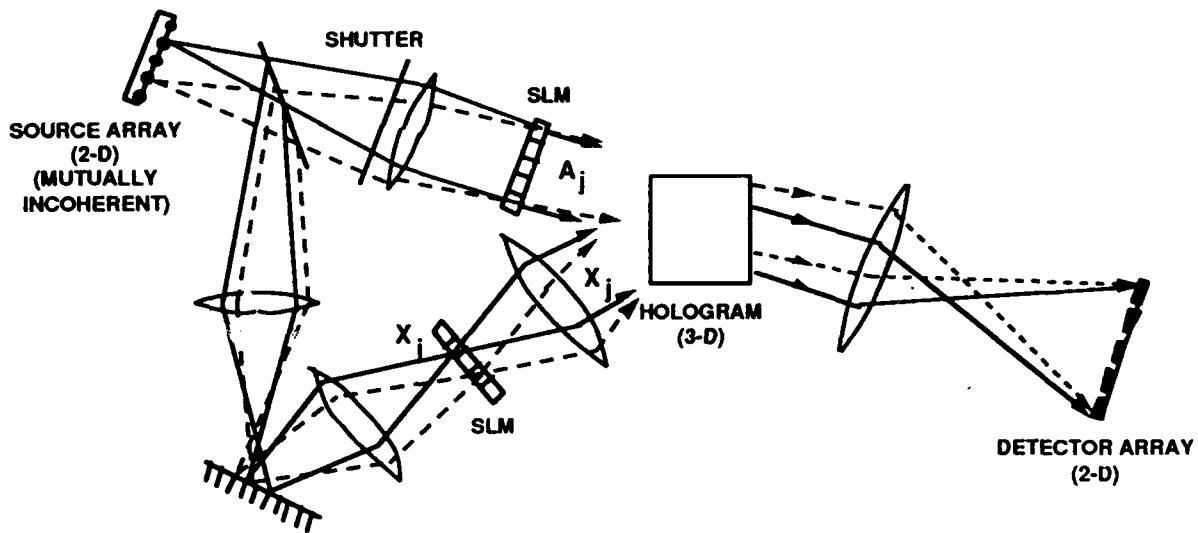


Figure 2. Interconnection system of the photonic architecture, showing incoherent/coherent double angular multiplexing geometry. The dashed beam corresponds to a different source than the solid beam. The two beams are mutually incoherent and are incident on the upper SLM at different angles.

interconnection weighted inputs; optimization of the overall interconnection throughput efficiency; and minimization of recording nonidealities introduced by the use of photorefractive media for grating formation. The reconstruction accuracy, residual crosstalk, and overall throughput efficiency of such an  $N^2$ - $N^2$  interconnection in the case of a linear holographic medium has been evaluated by numerical simulation using the optical beam propagation method (BPM)<sup>3</sup>. Results of a comparison between our proposed architecture and other candidate methods such as single-source fully coherent simultaneous recording show that our proposed method exhibits dramatically reduced crosstalk as well as significantly increased optical throughput. In addition, experiments in photorefractive and fixed holographic materials that we have performed to date support the viability of this approach to the implementation of highly multiplexed photonic neural network interconnections.

This photonic architecture potentially scales to  $10^{10}$  interconnections in one cascadable module, incorporating  $10^5$  neuron units running at (for the computing phases) 10-50 MHz, with fully parallel ( $10^5$  lines) input/output. Our technology development program has been directed toward the rapid realization of these goals, and our results have indicated that these goals are feasible.

### 2.3 Photonic Components

In order to assure system functionality, the requisite components incorporated in such an architecture must be designed from a synthetic viewpoint, such that their resulting performance parameters and characteristics are mutually compatible. In this section, the required features, design constraints, and achievements in the implementation of each essential component in the architecture are summarized.

The detection, amplification, neuron unit response, and modulation functions required in both the neuron unit and training planes are incorporated in multifunction spatial light modulators (SLM's). A dual rail differential approach is used because of its inherent capability to accommodate both bipolar inputs and outputs. The simpler case of unipolar outputs and bipolar inputs, also common in

neural network models, represents a subset of our fully bipolar design and requires even less chip area. The dual rail approach requires the hybrid or monolithic integration of two detectors, appropriate amplification and control circuitry, and two modulators within each SLM pixel. A primary goal has been to develop analog electronic circuitry that is process compatible with both detector and modulator requirements, and at the same time utilizes minimum real estate.

Our design incorporates silicon technology for detection and analog electronic neuron unit processing, and GaAs-based strained-layer multiple quantum well technology for modulation. As shown in Fig. 3, by flip-chip bonding two such chips face-to-face and operating at a wavelength for which the GaAs semi-insulating substrate is transparent, electrical contact between the two chips is provided at each pixel without either through-substrate vias or significant losses in chip area, bandwidth, or reliability. The design specification has used  $100 \mu\text{m} \times 100 \mu\text{m}$  pixels, divided into separate regions for the detection, modulation, and electronic circuitry.

Fabrication of  $6 \times 6$  arrays in silicon have shown sigmoidal response with full dual rail input and output capability, excellent uniformity over the array, and bandwidths of each pixel on the order of 10 MHz. Fabrication of InGaAs/GaAs Fabry-Perot cavity multiple quantum well reflective modulators in 2-D arrays have yielded a contrast ratio of 66:1 with a dynamic range of 30% for conventional cavity geometries<sup>4</sup>, and a contrast ratio of 30:1 with a dynamic range of 50% over a 2-D array in an inverted cavity geometry. This inverted cavity geometry<sup>5</sup> uses an *external* high-reflectivity cavity mirror, so that the modulator can be read out through the GaAs substrate, as required for flip-chip bonding to the silicon chip.

The two-dimensional array of sources must incorporate constant amplitude, individually coherent but mutually incoherent source elements. To assure mutual incoherence, the mutual coherence times between any pair of individual sources need only be less than the grating recording time constant of the holographic interconnection and the time constant of the neural detection circuitry. We give two examples of such source arrays. Two-dimensional arrays of surface emitting laser diodes have been fabricated by researchers at AT&T and Bellcore<sup>6</sup>, at a density of  $2 \times 10^6$  lasers per  $\text{cm}^2$ . Such lasers will be primarily mutually incoherent by the

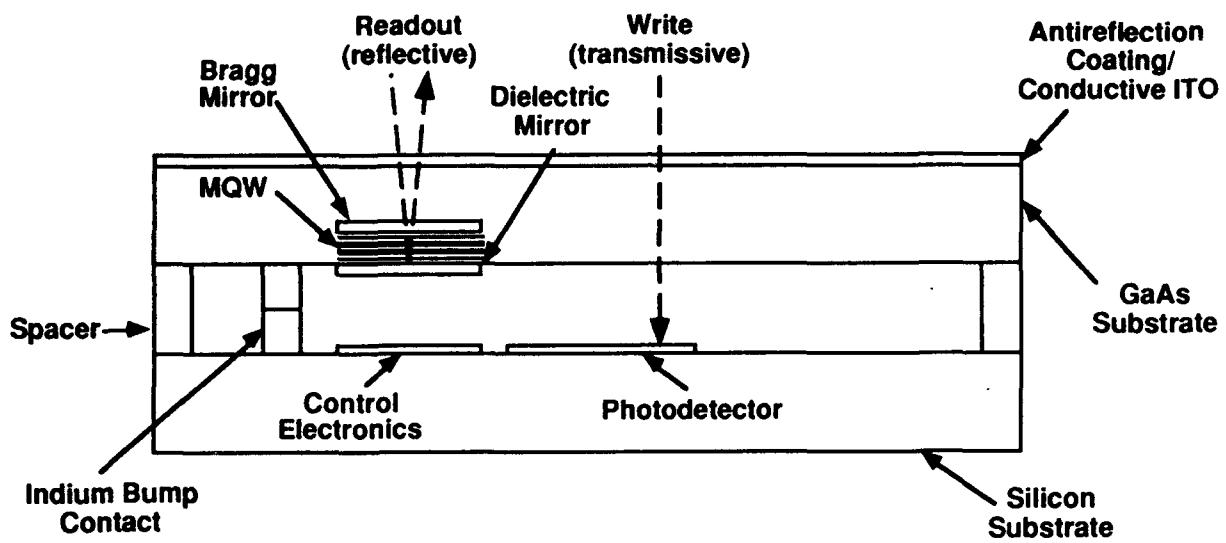


Figure 3. Cross section of one pixel of spatial light modulator structure, showing silicon detector and silicon analog electronic circuitry flip-chip bonded to GaAs-based reflection modulator. Designed wavelength of operation is approximately 950 nm.

very nature of nonuniformities inherent in the fabrication process. For laboratory demonstrations in the visible, we have fabricated acousto-optic source-array systems, each of which generates a set of beams from one coherent laser source. Each beam of the set of beams is Doppler shifted by a different amount, yielding effective mutual incoherence. This acousto-optic technique has been used successfully to record a set of holograms in photorefractive materials using the incoherent/coherent double angular multiplexing technique described above.

In conclusion, this research and development program encompasses the interdisciplinary areas of neural models, photonic architectures, photonic components, and photonic materials in order to develop a consistent technology for implementation of large-scale densely interconnected neural networks. Our use of an array of mutually incoherent sources provides for a novel incoherent/coherent double angular multiplexing technique for low crosstalk, high throughput interconnections; hybrid flip-chip bonded SLM's also provide for high speed neuron unit functionality in a near-term integrable package. We have also developed neural network models to ensure the self-consistency of the dual rail signal representation for neuron units and the corresponding interconnection functions. Our future plans include demonstration of the hybrid flip-chip bonded SLM that implements neuron unit arrays, and a system-level demonstration of neural network operation using this novel photonic architecture.

#### 2.4 References

1. Special Issue on Neural Networks, *Appl. Opt.*, Vol. 26, No. 23 (1 Dec. 1987).
2. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic Implementations of Neural Networks," Chap. 9 in B. Kosko, Ed., *Neural Networks for Signal Processing* (Prentice-Hall, Inc., Englewood Cliffs, NJ, 1992).
3. P. Asthana, G. P. Nordin, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of weighted fan-out/fan-in volume holographic optical interconnections," to appear in *Applied Optics*, special issue on neural networks (1993).

4. K. Hu, L. Chen, A. Madhukar, P. Chen, K. C. Rajkumar, K. Kaviani, Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., "High contrast ratio asymmetric Fabry-Perot reflection light modulator based on GaAs/InGaAs multiple quantum wells," *Appl. Phys. Lett.*, Vol. 59, No. 9, pp. 1108-1110 (1991).
5. K. Hu, L. Chen, A. Madhukar, P. Chen, C. Kyriakakis, Z. Karim, and A. R. Tanguay, Jr., "Inverted cavity GaAs/InGaAs asymmetric Fabry-Perot reflection modulator," *Appl. Phys. Lett.*, Vol. 59, No. 14, pp. 1664-1666 (1991).
6. J. L. Jewell, J. P. Harbison, A. Scherer, Y. H. Lee, and L. T. Florez, "Vertical-Cavity Surface-Emitting Lasers: Design, Growth, Fabrication, Characterization," *IEEE J. Quantum Electronics*, Vol. 27, No. 6, pp. 1332-1346 (June 1991).

### **3. List of Accomplishments**

**3.1 Development of a multilayer learning algorithm that exhibits complete bipolar functionality but can be implemented with only nonnegative signals.** This approach permits implementation of photonic neural networks that represent signals as (nonnegative) intensity levels and weights as (nonnegative) diffraction efficiencies, and that utilize dual channel inputs and outputs for each neuron unit. In our simulations (on 2-bit and 3-bit parity, and invariant TC recognition problems) this back-propagation style learning algorithm has performed comparably to, or better than, conventional back-propagation learning.

**3.2 Development, refinement, and generalization of the photonic neural network architecture.** Design of the generic architecture has been completed, and variants of the architecture for different neural network models, including multilayer back propagation, have been designed. The architecture is modular, permits fully parallel input/output, has a unique interconnection that yields crosstalk-free performance, and permits rapid copying of the interconnection pattern.

**3.3 Verification and characterization of incoherent/coherent holographic (fan-in/fan-out) interconnection systems via experiment and simulation.** Simulation at a variety of interconnection sizes, including 10 inputs to 10 outputs, have shown our novel (incoherent/coherent, double angularly multiplexed) holographic interconnection to exhibit high fidelity, relatively high diffraction efficiency, and extremely low crosstalk. In addition, experiments in photorefractive bismuth silicon oxide (2 inputs to 4096 outputs) and photorefractive lithium niobate (15 inputs to 16,384 outputs) yield further evidence that the underlying concepts are sound. This interconnection is a key element in our photonic neural network architecture.

**3.4 Design and fabrication of 2-D arrays of detectors and analog circuitry in silicon for neuron unit input and processing.** Six by six arrays of dual channel, sigmoidal function neuron units have been demonstrated using analog silicon CMOS (15 transistors within 2500 square microns per neuron unit), and exhibit bandwidths of 14 MHz (small signal) and 4 MHz (large signal). Measured output

voltages of 12 volts are sufficient for driving the GaAs-based modulators. Dual channel neuron unit circuitry incorporating analog sample and hold memory has also been designed and fabricated. For the neuron unit inputs, a variety of silicon photodetectors have been fabricated and tested, and used in cascade with the analog electronic sigmoidal function.

**3.5 Design and fabrication of 2-D arrays of optical modulators in GaAs/InGaAs for neuron unit output.** Multiple quantum well optical modulator arrays have been demonstrated in two configurations: an asymmetric cavity Fabry-Perot (exhibiting 66:1 contrast ratio and 30% dynamic range), and an inverted asymmetric cavity Fabry-Perot (exhibiting 30:1 contrast ratio and 50% dynamic range). Drive voltages, at 12 volts, are compatible with the silicon chip electronics. The 950 nm wavelength operating regime of these high bandwidth modulators provides compatibility with surface emitting semiconductor laser diode arrays, and due to substrate transparency permits direct flip-chip bump-contact bonding with the silicon chip.

**3.6 Generation of an array of self coherent but mutually incoherent sources.** An acousto-optic deflector based system has been designed, built, and demonstrated, that generates a linear array of 15 sources from a single laser source. A second system that can generate a 2-D array of up to 128 X 128 sources has been designed. In the 1 X 15 system, self coherence and mutual incoherence have been demonstrated by the simultaneous writing and individual readout of 15 independent gratings in photorefractive lithium niobate. This source array generator can be used with visible as well as infrared laser sources, and is useful for providing the mutually incoherent beams required in the photonic neural network architecture.

**3.7 Demonstration of recording and high-speed parallel readout of 2-D images on an optical disk, with potential applications to presentation of training patterns to a neural network during learning.** Recording and fully parallel readout of gray-level 512 x 512 images stored on an optical disk has been demonstrated. A novel differential interference contrast readout technique permits direct, optically efficient readout yielding high (50:1) contrast ratio images. This "optical disk spatial light

"modulator" can be used, for example, to provide a high frame-rate input to a photonic neural network.

#### **4. Publications and Patents to Date Resulting from this Grant**

##### **Publications**

1. B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay, Jr., "Photonic architecture for neural nets using incoherent/coherent holographic interconnections," *Proc. OC'90*, International Commission for Optics, Kobe, Japan, April 1990.
2. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., S. Piazzolla, B. K. Jenkins, and A. Madhukar, "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," *Proc. OC'90*, International Commission for Optics, Kobe, Japan, April 1990.
3. B. K. Jenkins, A. R. Tanguay, Jr., S. Piazzolla, G. C. Petrisor, and P. Asthana, "Photonic neural network architecture based on incoherent/coherent holographic interconnections," *Technical Digest*, Optical Society of America Annual Meeting, Boston, MA, Nov. 1990.
4. G. C. Petrisor, B. K. Jenkins, H. Chin, and A. R. Tanguay, Jr., "Dual function adaptive neural networks for photonic implementation," *Technical Digest*, Optical Society of America Annual Meeting, Boston, MA, Nov. 1990.
5. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., S. Piazzolla, B. K. Jenkins, "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," *Technical Digest*, Optical Society of America Annual Meeting, Boston, MA, Nov. 1990.
6. P. Asthana, G. Nordin, S. Piazzolla, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of interchannel crosstalk and throughput efficiency in highly multiplexed fan-out/fan-in holographic interconnections," *Technical Digest*, Optical Society of America Annual Meeting, Boston, MA, Nov. 1990.

7. C. Kyriakakis, Z. Karim, J. H. Rilum, J. J. Jung, A. R. Tanguay, Jr., and A. Madhukar, "Fundamental and technological limitations of asymmetric cavity MQW InGaAs/GaAs spatial light modulators," *Technical Digest*, OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada (1990).
8. Li Chen , K. C. Rajkumar, and A. Madhukar , *Appl. Phys. Lett.*, vol. 57, 2478(1990)
9. Li Chen , A. Madhukar, K. C. Rajkumar, K. Z. Hu, and J. J. Jung, Proceedings of the 20th International Conf. on Physics of Semiconductors (Aug. 6-10, 1990 Thessaloniki, Greece), World Scientific Publishing, Singapore (1990).
10. K. Z. Hu, L. Chen, A. Madhukar, P. Chen, K.C. Rajkumar, K. Kaviani, Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., "High Contrast Ratio Asymmetric Fabry-Perot Reflection Light Modulator Based on GaAs/InGaAs Multiple Quantum Wells", *Appl. Phys. Lett.*, 59(9), 1108-1110, (1991).
11. K. Z. Hu, L. Chen, A. Madhukar, P. Chen, C. Kyriakakis, Z. Karim, and A. R. Tanguay, Jr., "An Inverted Cavity GaAs/InGaAs Asymmetric Fabry-Perot Reflection Modulator", *Appl. Phys. Lett.*, 59(14), 1664-1666, (1991).
12. B. K. Jenkins, "Photonic interconnections for neural network implementations," invited paper, *Gordon Research Conference on Holography an Optical Information Processing*, Plymouth, New Hampshire (June 1991).
13. B. K. Jenkins and C. H. Wang, "Use of optics in neural vision models," *Technical Digest*, Annual Meeting of the Optical Society of America, paper MV1, San Jose, Calif. (1991).
14. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic Implementations of Neural Networks", Chapter 9 in *Neural Networks For Signal Processing*, B. Kosko, Ed., Prentice Hall, Englewood Cliffs, New Jersey, (1992).

15. Praveen Asthana, Edward Herbulock, Zaheed Karim, Chris Kyriakakis, Gregory P. Nordin, and Armand R. Tanguay, Jr., "Electrooptic Materials Requirements for Optical Information Processing and Computing", *Materials Research Society Proceedings*, Vol. 228, 1992, (in press).
16. Li Chen, K. Z. Hu, K.C. Rajkumar, S. Guha, R. Kapre and A. Madhukar, "Strained InGaAs/GaAs Multiple Quantum Wells Grown on Planar and Pre-Patterned GaAs(100) Substrates Via Molecular Beam Epitaxy: Applications to Light Modulators and Detectors", MRS Proceedings 228, 213 (1991)
17. K. Z. Hu, L. Chen, R. Kapre, K.C. Rajkumar, A. Madhukar, Z. Karim, C. Kyriakakis and A. Tanguay, Jr., "The Growth Performance of Strained InGaAs/GaAs Multiple Quantum Well Based Asymmetric Fabry-Perot Reflection Modulators", IEEE/LEOS Topical Meeting on "Epitaxial Materials and in Situ Processing for Optoelectronic Devices", July 29-31, 1991, Newport Beach, CA (pg.18).
18. L. Chen, K.Z. Hu, R.M. Kapre, W. Chen, and A. Madhukar, "High Contrast Optically Bistable Optoelectronic Switches Based on InGaAs/GaAs(100) Conventional and Inverted Asymmetric Fabry-Perot Modulators Grown via Molecular Beam Epitaxy", JVST B 10, 1014 (1992).
19. D. H. Rich, K.C. Rajkumar, Li Chen, A. Madhukar and F.J. Grunthaner, "Near-infrared cathodoluminescence study," J. Vac. Sci. Technol. 10, 1965 (1992).
20. D.H. Rich, K.C. Rajkumar, Li Chen, A. Madhukar and F.J. Grunthaner, "Near-infrared cathodoluminescence imaging of defect distributions in In<sub>0.2</sub>Ga<sub>0.8</sub>As/GaAs multiple quantum wells grown on prepatterned GaAs," Appl. Phys. Lett. 61, 222 (1992)
21. K. Hu, L. Chen, K. Kaviani, Q. Xie, K.C. Rajkumar, W. Chen, R.F. Cartland, P. Chen, and A. Madhukar, ICPS'92, Beijing, August 10-14, 1992.

22. B.K. Jenkins, A.R. Tanguay Jr., and A. Madhukar, "Photonic Technology for Densely-Interconnected Neural Networks," Government Microcircuit Applications Conference, Las Vegas, Nevada, pp. 563-566, November 9-12, 1992.
23. P. Asthana, G. P. Nordin, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of Weighted Fan-out/Fan-in Volume Holographic Interconnections," Applied Optics, Vol. 32, pp. 1441-1469, 10 March 1993.
24. G. P. Nordin, P. Asthana, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of Weighted Fan-out-Fan-in Volume Holographic Interconnections," Topical Meeting on Diffractive Optics: Design, Fabrication, and Applications, 1992 Technical Digest Series, Vol. 9, pp. 165-167 (Optical Society of America, Washington, D. C. 1992).
25. B. K. Jenkins and C. H. Wang, "Use of Optics in Neural Vision Models," Annual Meeting of the Optical Society of America, invited paper MV1, San Jose, Calif., November 4-9, 1991.
26. G. C. Petrisor, S. Piazzolla, G. Nordin, P. Asthana, E. Herbulock, B. K. Jenkins, and A. R. Tanguay, Jr., "Experimental Evaluation of Incoherent/Coherent Volume Holographic Interconnection Systems, Annual Meeting of the Optical Society of America, Albuquerque, New Mexico, September 20-25, 1992, paper FBB3
27. B. K. Jenkins, "Photonic Neural Networks with Applications to Vision," IEEE/SICE International Workshop on Emerging Technologies for Factory Automation, Palm Cove-Cairns, Australia, 17-19 August 1992.

## Patents

1. B. K. Jenkins and A. R. Tanguay, Jr., "Incoherent/coherent multiplexed holographic recording for photonic interconnections and holographic optical elements," U. S. Patent No. 5,121,231, June 9, 1992.
2. B. K. Jenkins and A. R. Tanguay, Jr., "Spatial light modulators for incoherent/coherent multiplexed holographic recording and readout," USC-2254A, divisional, U. S. P. T. O. filing date April 6, 1990, patent pending.
3. B. K. Jenkins and A. R. Tanguay, Jr., "Incoherent/coherent source array for multiplexed holographic recording and readout," USC-2254B, divisional, U. S. P. T. O. filing date April 6, 1990, patent pending.
4. B. K. Jenkins and A. R. Tanguay, Jr., "Incoherent/coherent double angularly multiplexed volume holographic optical elements," USC-2254C, continuation in part, U. S. P. T. O. filing date June 8, 1992, patent pending.

**5. Copies of Selected Publications for Further Information**

## Analysis of Weighted Fan-out/Fan-in Volume Holographic Interconnections

**Gregory P. Nordin**

University of Southern California  
University Park MC-0483  
Los Angeles, CA 90089-0483

**Praveen Asthana**

IBM Corporation  
9000 S. Rita Rd.  
Tucson, AZ 85715

**Armand R. Tanguay, Jr. and B. Keith Jenkins**

University of Southern California  
University Park MC-0483  
Los Angeles, CA 90089-0483

### Summary

Volume holography has often been proposed as an interconnection technology for photonic neural network implementations [1] because it appears to potentially meet the critical requirements of providing both large numbers of interconnections *and* a weighted fan-out/fan-in topology. The feasibility of using volume holograms for large-scale weighted fan-out/fan-in interconnection applications depends in part on the *fidelity* with which the interconnection weights can be implemented, and on the *throughput* that can be achieved in the volume holographic interconnection system. In addition, an important implementation issue is the *number of exposures* required to record a particular interconnection configuration because this can impact the total writing time and required exposure schedule.

In many neural network algorithms, the interconnections between any two layers can be represented as  $\rho = Wx$  in which each component of the vector  $x$  represents the output of a neuron unit in the input layer,  $W$  is the interconnection matrix, and each component of the vector  $\rho$  is the input to a neuron unit in the output layer. A typically proposed volume holographic interconnection architecture that performs this interconnection function [1] (hereafter referred to as a "conventional" architecture) is shown in Fig. 1. Each interconnection between a pixel in the input plane and a pixel in the output plane (not shown) is implemented as a single diffraction grating in the holographic recording medium. The weight of the interconnection is related to the relative diffraction efficiency of its associated grating.

For adaptive neural networks in which a learning algorithm is implemented, the weight matrix is typically formed by successively presenting training patterns (consisting of an input image at the input plane and its associated training image at the training plane) to the system until the appropriate error criterion is minimized. In a conventional architecture, there are three methods to record the desired interconnection gratings during the presentation of a single training pattern to the system. In the first method, all of the pixels in the input and training planes are turned on simultaneously. Since the light from each pixel is mutually coherent, the desired interconnection gratings are formed. In addition, undesired gratings are created that form *intraplanar* connections amongst the pixels of both the input plane and the training plane. As is generally recognized, these additional coupling paths introduce a serious source of crosstalk into the interconnection system. By recording the desired gratings in a page-wise sequential (*i.e.*, a single pixel in the input plane simultaneously with all of the training plane pixels) or fully sequential (*i.e.*, a single input pixel with a single training pixel) manner, some or all of these coherent crosstalk gratings may be avoided. The cost, however, is to require additional exposures compared to the simultaneous recording method. For an interconnection system connecting  $N$  input nodes to  $N$  output nodes (an "N-N" interconnection), the simultaneous recording method requires one exposure per training pattern whereas the page-wise sequential and fully sequential methods require  $N$  and  $N^2$  exposures, respectively. For large numbers of interconnections (the situation for which holographic interconnections are presumably attractive), this could result in impractically long training sessions.

We quantitatively examined the reconstruction fidelity and throughput of the conventional architecture by numerically modeling a 10-10 interconnection system using the optical beam propagation method. A linear holographic recording medium was assumed with unlimited modulation range. Ten training patterns of random intensities were used to generate the (nearly independent) weight matrix recorded in the holographic medium. For this choice of weights, the RMS error of the diffracted outputs [2] for a conventional architecture is shown in Fig. 2a for simultaneous (with  $R=1$ , in which  $R$  is the beamsplitter ratio in Fig. 1), page-wise sequential ( $R=100$ ), and fully sequential recording ( $R=1$ ). The horizontal axis represents the grating strength (in radians) of the largest interconnection grating recorded in the medium. The interconnection is readout with a set of mutually coherent random intensity beams. The optical throughput is shown in Fig. 2b. As expected, the RMS error of the reconstructed outputs is large for simultaneous recording. There is little improvement in the error in going to page-wise sequential recording using a beamsplitter ratio of 100. This can be understood as follows. During the recording of a single training pattern, the crosstalk gratings connecting the pixels within the training plane are exposed  $N$  times. The desired gratings, however, each receive only a single exposure. When the undesired crosstalk gratings are eliminated using fully sequential recording, the RMS error drops dramatically such that it is still relatively low at a peak throughput of over 85%.

An alternative volume holographic interconnection architecture has recently been proposed [3] that requires only one exposure per training pattern and yet minimizes undesired crosstalk gratings. This architecture includes an optical source array in which each source is individually coherent, but mutually incoherent with all other sources. In Fig. 3 we show a modification of this architecture that for implementation in certain photorefractive materials does not suffer from an inherent throughput loss due to multiple beam superposition. Using lenses  $L_1$  and  $L_2$  (see Fig. 3) a set of subholograms (which may partially overlap) are recorded in the holographic medium. Each subhologram consists of a 1- $N$  fan-out connecting a single input pixel to every output pixel (with the aid of an imaging lens following the holographic medium [not shown]).

Numerical modeling results for this multiple source incoherent/coherent architecture (for  $R = 100$ ) are shown in Fig. 4 (in which non-overlapping subholograms are assumed). The weight matrix and random inputs are the same as used above for the conventional architecture. The RMS error of the diffracted outputs is comparable to the results obtained for fully sequential recording in a conventional architecture, and yet the multiple source incoherent/coherent architecture requires only a single recording per training pair.

In addition to requiring fewer exposures to achieve the same level of fidelity and throughput as obtained for fully sequential recording in a conventional architecture, the multiple source incoherent/coherent architecture offers several further advantages. First, the weighted interconnections stored in the volume holographic medium can be copied to another volume holographic medium in a one-step process (which is useful, for example, for transferring a set of learned weights from a dynamic to a permanent holographic medium) [4]. Second, since the source array shown in Fig. 3 consists of optical sources that are individually coherent but mutually incoherent, the summation at the output nodes is performed by the fan-in of mutually incoherent beams. This eliminates the requirement for strict positional stability of the optical components during readout (which is required in a conventional architecture for readout with mutually coherent beams). No degradation in throughput (see Fig. 4b) is suffered with the use of mutually incoherent beams because an angular fan-in is used at each output node.

- [1] D. Psaltis, D. J. Brady, and K. Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.* 27, p. 1752-1758.
- [2] C. Slinger, "Analysis of N-to-N Volume-Holographic Neural Interconnect," *J. Opt. Soc. Am. A* 8, p. 1074-1081.
- [3] B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay Jr., "Photonic Architecture for Neural Nets Using Incoherent/Coherent Holographic Interconnections," *Proceedings of OC'90*, Kobe, Japan (1990).
- [4] S. Piazzolla, B. K. Jenkins, and A. R. Tanguay, Jr., "Single Step Copying Process For Multiplexed Volume Holograms," submitted to *Optics Letters*.

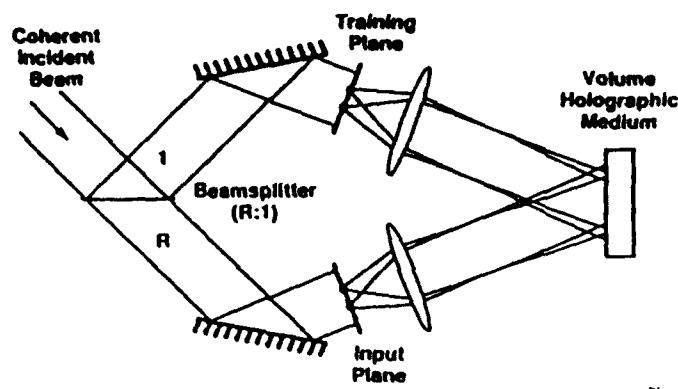


Figure 1

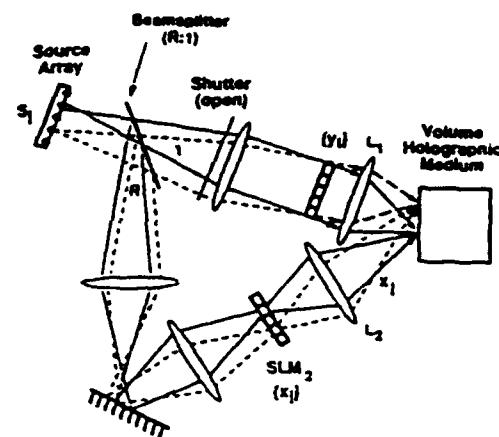


Figure 3

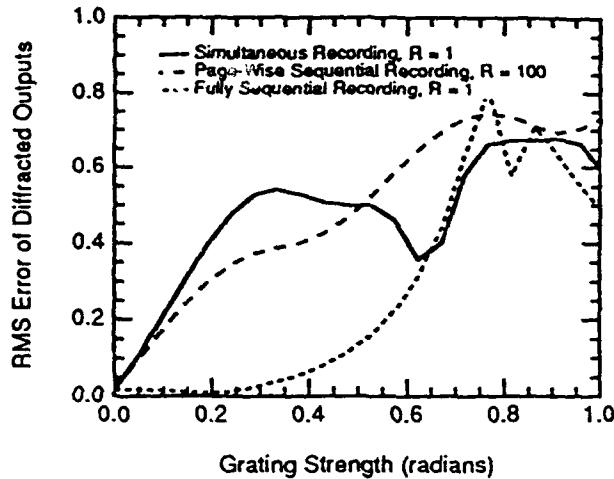


Figure 2a

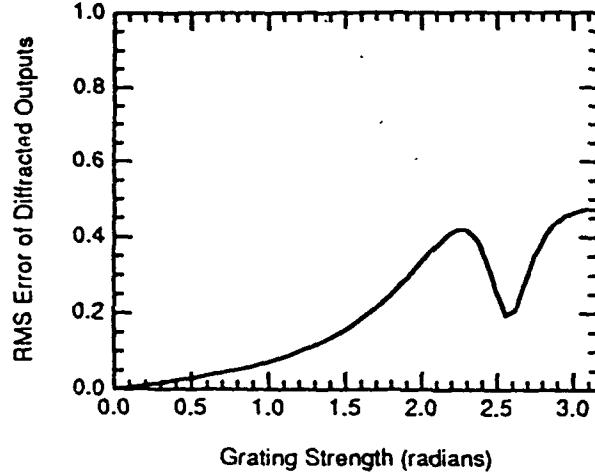


Figure 4a

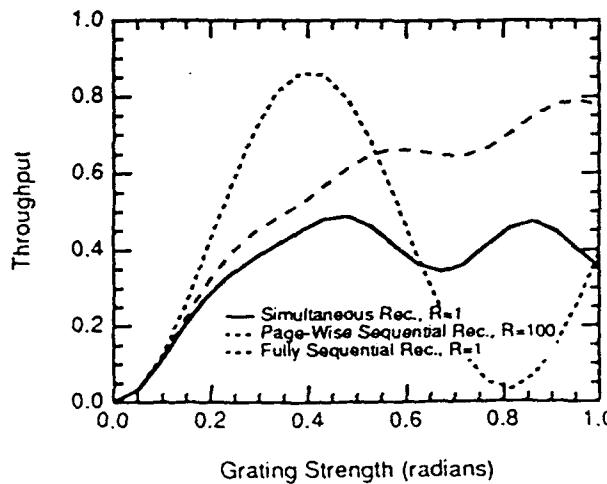


Figure 2b

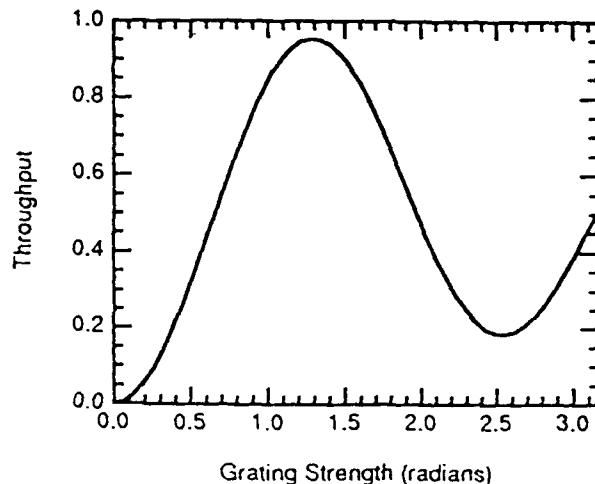


Figure 4b

# STRAINED InGaAs/GaAs MULTIPLE QUANTUM WELLS GROWN ON PLANAR AND PRE-PATTERNEDE GaAs(100) SUBSTRATES VIA MOLECULAR BEAM EPITAXY: APPLICATIONS TO LIGHT MODULATORS AND DETECTORS

Li Chen, Kezhong Hu, K. C. Rajkumar, S. Guha\*, R. Kapre and A. Madhukar  
Photonic Materials and Devices Laboratory, University of Southern California,  
Los Angeles, CA 90089-0241

## ABSTRACT

We report the realization of high quality strained InGaAs/GaAs multiple quantum wells (MQW) grown on planar GaAs (100) substrates through optimization of molecular beam epitaxial (MBE) growth conditions and structure. Such MQWs containing ~11% In have lead to the realization of an asymmetric Fabry-Perot (ASFP) reflection modulator with a room temperature contrast ratio of 66:1 and an on-state reflectivity of 30%. For In composition  $\geq 0.2$ , the improved optical quality for very thick ( $>2\mu\text{m}$ ) InGaAs/GaAs MQWs grown on pre-patterned substrates is demonstrated via transmission electron microscopy (TEM) and micro-absorption measurements.

Strained InGaAs/AlGaAs layered structures grown on GaAs(100) substrates have attracted considerable attention for their application in vertical cavity surface emitting lasers [1], resonant tunneling diodes (RTD) [2], and light modulators [3,4]. In the transmission geometry, such modulators can take advantage of the transparent nature of the GaAs substrate so that, unlike the GaAs/AlGaAs based modulators, there is no need for a patterned removal of the GaAs substrate. Reflection mode ASFP modulators, when combined with RTDs and Si field effect transistors, provide optical digital switches with high fan-out useful for digital optical computing and communication networks [5]. In the reflection geometry, *inverted* reflection modulators [6] can also be made in which light passes in and out through the substrate so that a Si-chip containing control electronics can be flip-chip bonded to the modulator bearing GaAs chip [7]. When combined with arrays of strained InGaAs/GaAs lasers, detectors and Si-control electronics, analog optical neuron units can be realized.

Unlike strained InGaAs/AlGaAs quantum well lasers and RTDs, which require relatively thin active layers, the major requirement for light modulators is to grow the strained active region to thicknesses ( $\sim 1.5\mu\text{m}$ ) necessary to obtain reasonable optical interaction path length. Through control of the growth kinetics via reflection high-energy electron diffraction (RHEED) pattern and intensity dynamics [8] in molecular beam epitaxial (MBE) growth, we have realized sharp exciton linewidths in such multiple quantum wells and thus achieved a typical absorption modulation per well, ( $\Delta\alpha_{d_w}$ ) of  $\sim 0.004$  in the exciton tail region at relatively low bias ( $\sim 80$  kV/cm) [4]. Placing these strained MQWs in an ASFP configuration involving integrated Bragg mirrors it should thus be possible to obtain high contrast ratios. Contrast ratios of  $> 100:1$  have been demonstrated [9,10] in the AlGaAs/GaAs system but have been limited to

---

\* Present address: IBM T. J. Watson Research Center, Yorktown Height, NY 10598

~3:1 in the InGaAs/GaAs system [11]. The high contrast ratio in the ASFP configuration is obtained by modulating, say electrically, the absorption in the MQW so that at one of the voltages the primary reflected beam is canceled by the higher order reflected beams and a near zero reflectivity is achieved. Calculations indicate that with front and back mirror reflectivities of 0.68 and 0.99 respectively, it is possible to reach this low reflection state for  $\Delta\alpha \cdot d_w \cdot N$  (where  $N$ =number of wells) of ~ 0.15 when operating near 9500Å (In content ~ 12%). For higher In composition ( $\geq 20\%$ ), to circumvent the problem of dislocation generation at the needed thickness, we have applied the physical idea of defect reduction through growth on pre-patterned substrates [12,13]. Here we show that even when the growth on the planar part of the substrate exhibits high density of structural defects and no excitonic features, the MQWs grown on the pre-patterned part (both on top of the mesa and, surprisingly, in between mesas) still show nearly perfect layering and sharp excitonic features with linewidths comparable to those obtained on samples of lower In composition. Using this defect reduction technique, it is thus possible to grow such highly strained MQWs with integrated Bragg mirrors to realize high performance ASFP modulators operating near 9800Å wavelength made important by the developments in Er doped optical fibers and at 1.06μm wavelength of interest due to the availability of Nd:YAG lasers.

The structures were grown in our RIBER 32P MBE system on either Cr-doped semi-insulating or Si-doped n<sup>+</sup> GaAs (100) substrates. The temporal behavior of the RHEED pattern and the specular beam intensity was employed to obtain optimized growth conditions, ensure reproducible growth conditions, and to monitor accurately the group III incorporation rates.

Fig.1 shows the electromodulation behavior of samples RG900420 and RG900726 which consist of 50 period In<sub>0.11</sub>Ga<sub>0.89</sub>As(100Å)/GaAs(200Å) and 20 period In<sub>0.19</sub>Ga<sub>0.81</sub>As(80Å)/GaAs(145Å) MQWs, respectively, grown on non-patterned GaAs(100) substrates. For comparison, data for InGaAs/GaAs MQWs appearing in the literature [14-16] are also included. The product of the maximum change in absorption coefficient ( $\Delta\alpha$ ) and the well width ( $d_w$ ), which gives modulation per well, is plotted for  $\Delta\alpha$  in the exciton peak and tail regions as a function of the electric field with the built-in field accounted for. The room temperature exciton half widths at half maximum (HWHM) for RG900420 and RG900726 are 5.5meV and 7.5meV, respectively. Sample RG900420 clearly shows very large  $\Delta\alpha d_w$  of ~ 0.004 in the exciton tail region at a low electric field of ~ 60 kV/cm. Not shown in the figure, a sample RG900802 (comprised of 50 period In<sub>0.13</sub>Ga<sub>0.17</sub>As(100Å)/GaAs(125Å) ) gives a room temperature HWHM of 6.2meV and per well modulation of ~0.004 at a field of ~ 90kV/cm.

Given these MQW characteristics, ASFP reflection modulators were grown in which such MQWs are sandwich between high quality Al<sub>x</sub>Ga<sub>1-x</sub>As/Al<sub>y</sub>Ga<sub>1-y</sub>As Bragg mirrors. Our sample RG910115 consists of 17.5 periods of AlAs(784Å)/GaAs(654Å) as the bottom mirror with computed reflectivity  $R_b = 99.4\%$ , 3 periods for the top mirror of computed  $R_t = 68\%$ , and in

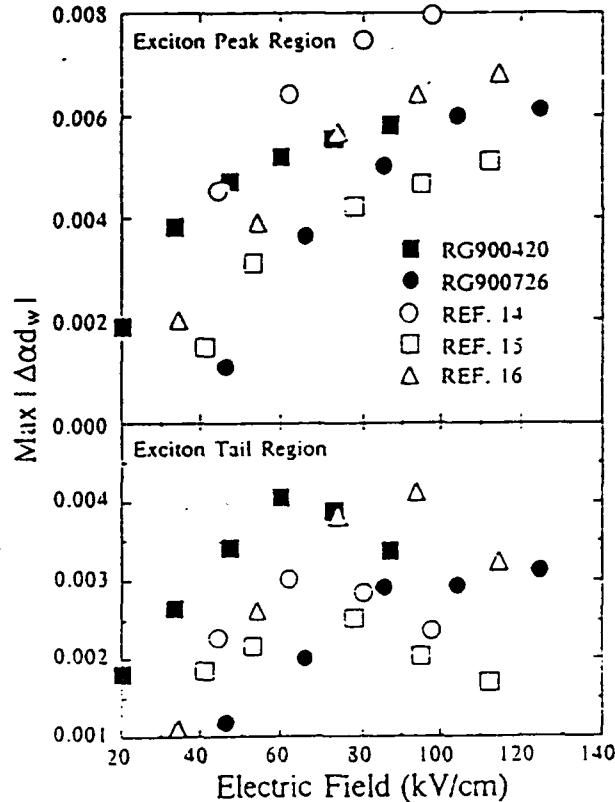


Fig. 1 Comparison of the product of maximum change in the absorption ( $\Delta\alpha$ ) and well width ( $d_w$ ) for USC sample RG900420 and RG900726 with others in the literature.

between a 50 period  $In_{0.11}Ga_{0.89}As(100\text{\AA})/GaAs(125\text{\AA})$  which is expected to have similar behavior as samples RG900420 and RG900802 discussed above. The bottom and top mirrors are  $n^+$  and  $p^+$  doped except for the last and the first period, respectively. The grown samples were patterned into  $280\mu\text{m}$  diameter test pixels and  $125\mu\text{m}$  diameter Cr-Au contact pads were deposited on top of the mesas to provide top electrical contact. The back contact is provided by the In used for mounting the GaAs substrate on the Mo block of the MBE system. The reflection spectra were taken in a normal incidence geometry with an automated Ti/Sapphire laser pumped by an  $\text{Ar}^+$  laser, a beam splitter cube, and two Si detectors. Fresh deposited Au was used for the reflectivity calibration.

Fig. 2 (a) shows the reflectivity of pixel (4, 31) of this sample at 0V, 8V and 16V reverse bias. Note that by about -8V applied bias the reflectivity at the zero bias F-P wavelength has dropped from ~19.2% to ~1.4% with a slight red shift in the FP position itself. The modulation effect is thus predominantly an electroabsorptive one in this regime. With further increase in the reverse bias, the reflectivity has dropped to ~0.46% at the blue-shifted F-P wavelength of  $9381\text{\AA}$  at -16V. The blue shift in the F-P wavelength is a consequence of electro-refraction setting in. The reflectivity contrast ratio (CR) and change ( $\Delta R$ ) between zero and -16V applied bias are shown as a function of the wavelength in Fig. 2, panels (b) and (c), respectively. A maximum contrast ratio of 66:1 with an attendant  $\Delta R$  of ~30% is seen at  $9381\text{\AA}$  corresponding to the modulator operation in the 'normally-on' condition. The details are reported in Ref. 17.

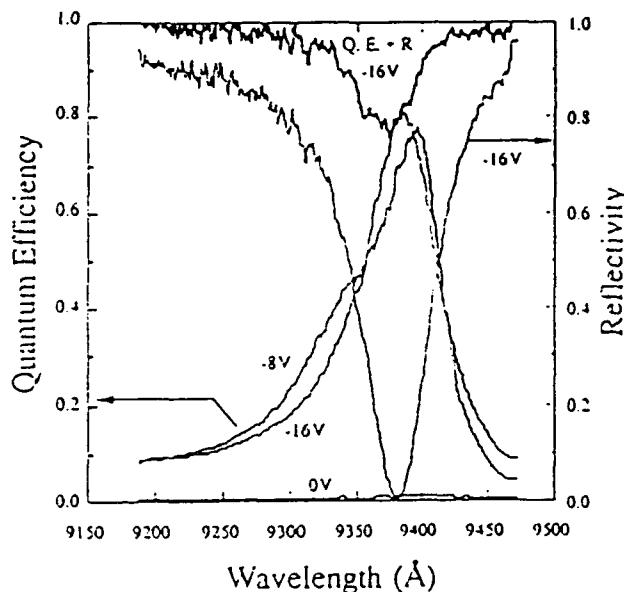


Fig. 3 Quantum efficiency (Q.E.) at 0V, -8V, and -16V biases and reflectivity (R) at -16V bias of pixel (4,31) as a function of wavelength. The thick curve on the top is the sum of Q.E. and R at -16V bias.

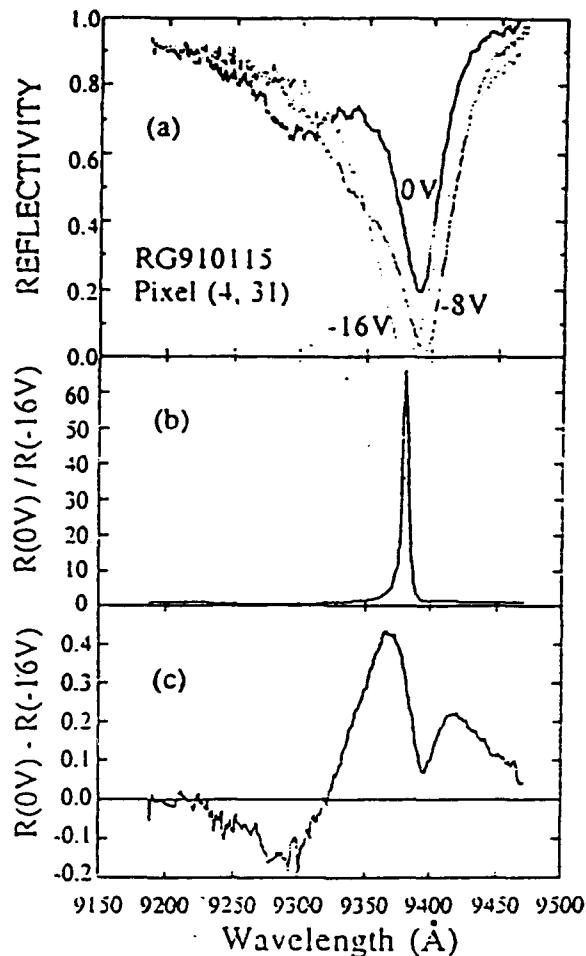


Fig. 2 Normalized reflectivity as a function of wavelength at different applied biases for pixel (4,31). The reflectivity change and contrast ratio between zero and -16V bias are shown in panels (b) and (c), respectively.

Since the ASFP is in a p-i-n configuration, such a modulator can also be used as a detector. Fig. 3 shows the quantum efficiency of the same pixel at 0V, -8V and -16V. The near zero quantum efficiency at zero bias is indicative of the nearly flat-band condition operative. This was a design goal achieved by no doping of the last period of the bottom  $n^+$  mirror and the first period of the top  $p^+$  mirror. At -8V the quantum efficiency at  $\lambda_{FP}$  is near 75%

and by -16V it has reached near 82%. The reflectivity spectrum at -16V is reproduced here to show that, as expected, the high quantum efficiency at  $\lambda_{FP}$  is a consequence of the increased absorption due to multiple reflection under the FP resonance condition. Without the FP cavity, the absorption depth,  $\alpha d$ , at the FP wavelength is estimated to be  $\sim 0.15$  and thus will give maximum quantum efficiency of  $\sim 14\%$  only. The sum of the reflectivity and quantum efficiency at -16V, shown in Fig. 3, is seen to be near unity away from the F-P resonance region but dips to near 80% at  $\lambda_{FP}$ . The loss is due to enhanced 'leakage' of the light through the bottom high reflectivity mirror due to the multiple in-phase transmissions at FP resonance. This leakage is extremely sensitive to even small variations in the reflectivity of the high reflectivity mirror. From the nearly 20% leakage we estimate the operative reflectivity at the -16V  $\lambda_{FP}$  to be 97.8%. This is to be compared with the 99.4% reflectivity design value computed at zero bias on the basis of the GaAs/AlAs quaterwave stack.

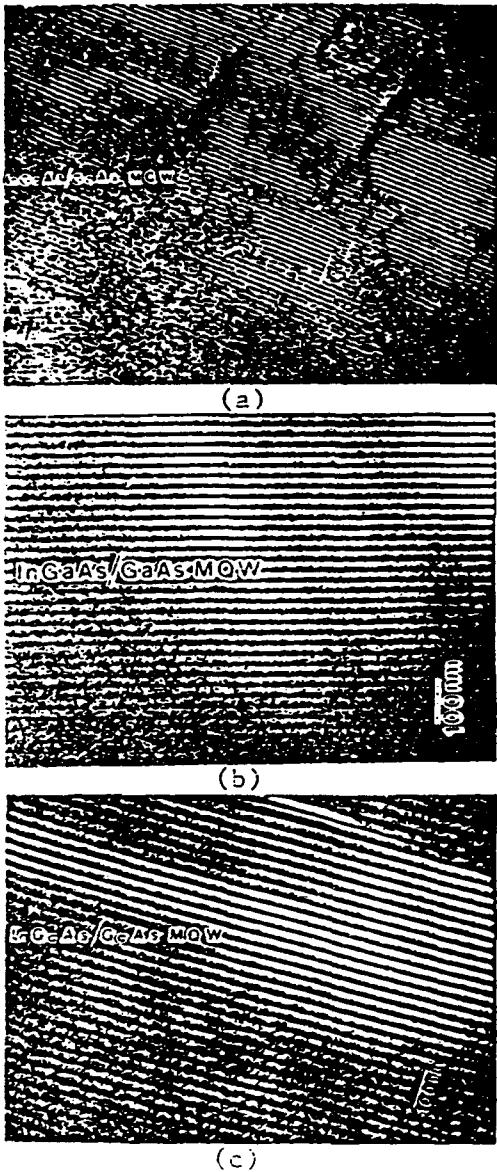


Fig.4 Cross-sectional TEM image contrast of a 100 period  $In_{0.2}Ga_{0.8}As$  (80 Å)/GaAs(160 Å) MQW (Sample RG891110) taken for the growth on the nonpatterned region (a) and in the patterned part between mesas (b) and on top of the mesa (c).

To improve the optical quality of the InGaAs/GaAs MQW with higher In content and large thickness, structures were grown on the GaAs (100) substrates partially patterned with mesas of various shapes and sizes. The substrate is first chemically cleaned using standard MBE cleaning procedures and then patterned using photolithography and wet chemical etching as reported in our earlier work [13]. An example is sample RG891110 which consists of a 100 period  $In_{0.2}Ga_{0.8}As$ (80Å)/GaAs(160Å) MQW of  $\sim 2.38 \mu m$  total thickness. In the patterned part, the mesa size is  $\sim 16 \mu m \times 18 \mu m$  and the pitch is  $40 \mu m$  in both directions. Fig.4(a) shows a cross sectional transmission electron microscope (XTEM) image contrast taken with the electron beam along the  $[1\bar{1}0]$  azimuth for a specimen prepared from the non-patterned part of the sample. A large number of structural defects, including threading dislocations, are seen to be present throughout the MQW structure. Figs. 4(b) and 4(c) show the corresponding image contrast taken (at higher magnification) in the patterned part between mesas and on top of the mesa, respectively. Remarkably perfect InGaAs (dark) and GaAs (light) layers are seen with hardly any evidence of structural defects at the level of the TEM resolution. However, towards the edges of the mesas, the layering was found not to be good within 2 to 3  $\mu m$  of the edges, even though no defects were visible using the customarily chosen  $[1\bar{1}0]$  azimuth. The presence of contiguous mesa top plane and the sidewall planes (typically  $(1\bar{1}1)$  with some high index facets such as  $\{3\bar{1}1\}/\{411\}$ ,  $\{1\bar{1},1,1\}$  etc.) gives rise to interfacial cation migration whose degree and nature is sensitive to the growth conditions, as we have previously demonstrated [13]. We therefore suspect that interfacial migration is in part, if not largely, responsible for the degradation of the sharpness of layering near the edges of the mesas.

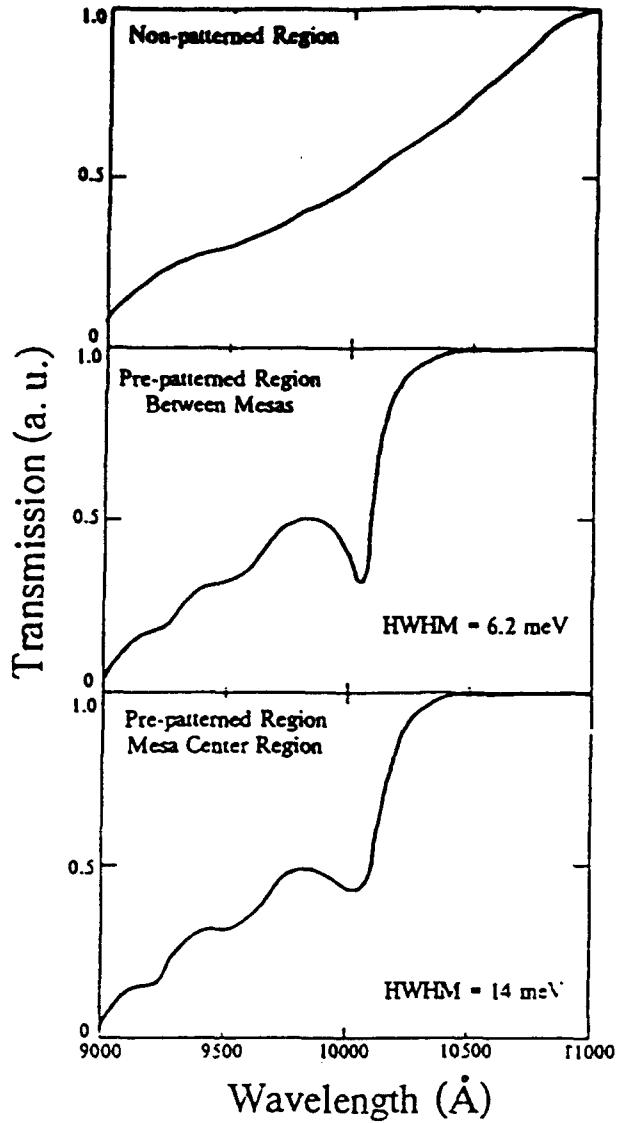


Fig.5 Optical transmission behavior of sample RG891110 for the MQW in the nonpatterned region (a) and patterned part between mesas (b) and on top of the mesa (c).

Fig. 5, panels (a), (b) and (c) show the corresponding room temperature transmission characteristics of this sample. The transmission spectra were obtained using a highly collimated beam sampled from a 50W tungsten halogen lamp, a 100x microscope objective lens which results in 2 to 3  $\mu\text{m}$  beam size on the sample. A SPEX 1704 monochromator and a  $\text{LN}_2$ -cooled Ge detector was used for detection. The spectra are normalized to the system response. The non-patterned region shows no excitonic features, consistent with the TEM findings (Fig. 4a). By contrast, the MQW growth on top of the mesas and in between the mesas shows not only the usual heavy hole (hh) to first confined electron (1e) excitonic transitions at  $\sim 1 \mu\text{m}$  wavelength, but even the light hole (lh) to 1e and second hh to second electron excitonic features are resolved, indicating the high quality of the MQW in these regions. This also is consistent with the TEM findings (figs. 4(c) and 4(b)). The HWHM of the exciton for the MQW in the region between the mesas and on top of the mesa are 6.2meV and 14meV, respectively. While a clear explanation of the sharper excitonic feature in the trench region requires further examination, it may be speculated that it is partially due to the interfacet migration since it seems that the high index planes move towards the center of the mesa more than they move outwards. The trench regions thus provide regions for realization of good light modulator pixels. The linewidth approaching the value obtained for lower In composition MQWs grown on non-patterned substrates and which, in the ASFP configuration, give very high contrast ratio and reasonable throughput indicates that such MQWs can offer good electromodulation behavior at wavelength near 1.06  $\mu\text{m}$ .

In conclusion, we have demonstrated that high quality strained InGaAs/GaAs ( $x \leq 0.15$ ) MQWs of large thicknesses ( $\leq 2 \mu\text{m}$ ) and with good electroabsorption characteristics can be grown individually and with integrated Bragg mirrors by optimizing the MBE growth condition. The resulting material has lead to the successful demonstration of InGaAs/GaAs strained ASFP modulators with room temperature contrast ratios as high as 66:1 while maintaining a high throughput of 30%. These modulators provide some technological advantages over the AlGaAs/GaAs modulators and can be integrated with other III-V devices as well as Si electronics to realize optical neuron units [6] and digital photonic switches [5]. The behavior of the latter, utilizing ASFP modulators and detectors realized as part of this work, is presented in the paper by Kapre et. al. in this volume. At high In compositions ( $\geq 20\%$ ), high quality and very thick ( $\sim 2 \mu\text{m}$ ) InGaAs/GaAs quantum wells have been achieved through growth on pre-patterned substrates and are expected to be useful for high contrast ratio modulators near 1  $\mu\text{m}$ .

This work was supported by the AFOSR, URI(AFOSR), DAPRA and ONR.

**References:**

1. For example, see J. L. Jewell, Y. H. Lee, A. Scherer, S. L. McCall, N. A. Olsson, J. P. Harbison and L. T. Florez, *Optical Engineering* 29, 211 (1990).
2. R. Kapre, A. Madhukar, K. Kaviani, S. Guha, and K. C. Rajkumar, *Appl. Phys. Lett.* 56, 922 (1990).
3. T. K. Woodward, Theodore Sizer, II, E. L. Sivco, and A. Y. Cho, *Appl. Phys. Lett.* 57, 548 (1990).
4. Li Chen, K. C. Rajkumar, and A. Madhukar, *Appl. Phys. Lett.* 57, 2478 (1990).
5. R. Kapre, Kezhong Hu, Li Chen, A. Madhukar and S. Guha, these Proceedings; see also, Li Chen, R. M. Kapre, Kezhong Hu, and A. Madhukar, *Appl. Phys. Lett.* (submitted).
6. Kezhong Hu, Li Chen, Anupam Madhukar, Ping Chen and Armand R. Tanguay, Jr., *Appl. Phys. Lett.* (submitted).
7. C. Kyriakakis, Z. Karim, J. H. Rilum, J. J. Jung, A. R. Tanguay, Jr., and A. Madhukar, OSA Topical Conference on Spatial Light Modulators and Applications, Incline Village, Nevada, Vol.14 of the 1990 OSA Technical Digest Series 14, 7-10, 1990.
8. P. Chen, J. Y. Kim, A. Madhukar, and N. M. Cho, *J. Vac. Sci. Technol.* B4, 890 (1986).
9. K-K Law, R. H. Yan, L. A. Coldren, and J. L. Merz, *Appl. Phys. Lett.* 57, 1345 (1990).
10. M. Whitehead, A. Rivers, G. Parry, J. S. Roberts and C. Button, *Electron. Lett.* 25, 984, (1989).
11. B. Pezeshki, D. Thomas and J. S. Harris Jr., *IEEE Photon. Technol. Lett.* 2, 807, 1990.
12. E. A. Fitzgerald, G. P. Watson, R. E. Proano, D. G. Ast, P. D. Kirchner, G. D. Pettit and J. M. Woodall, *J. Appl. Phys.* 65, 2220 (1989).
13. S. Guha, A. Madhukar, K. Kaviani, and R. Kapre, *J. Vac. Sci. Technol.* B8, 149 (1990). See also, S. Guha, A. Madhukar, and Li Chen, *Appl. Phys. Lett.* 56, 2304, 1990.
14. T. E. Van Eck, P. Chu, W. S. C. Chang, and H. H. Wieder, *Appl. Phys. Lett.* 49, 135 (1986).
15. W. Dobbelaere, S. Kalem, D. Huang, M. S. Unlu, and H. Morkoc, *Electron. Lett.* 24, 295 (1988).
16. W. D. Goodhue, B. E. Burke, B. F. Aull, and K. B. Nichols, *J. Vac. Sci. Technol.* A6, 2356 (1988).
17. Kezhong Hu, Li Chen, A. Madhukar, P. Chen, K. Kaviani, Z. Karim, C. Kyriakakis and A. R. Tanguay, Jr., *Appl. Phys. Lett.* (submitted).

# High contrast optically bistable optoelectronic switches based in InGaAs/GaAs(100) conventional and inverted asymmetric Fabry-Pérot modulators grown via molecular-beam epitaxy

Li Chen, Kezhong Hu, R. M. Kapre, W. Choi, P. Chen, and A. Madhukar  
Photonic Materials and Devices Laboratory, University of Southern California, Los Angeles,  
California 90089-0241

(Received 16 September 1991; accepted 29 October 1991)

High contrast ratio InGaAs based asymmetric Fabry-Pérot reflection modulators, realized in both conventional and inverted geometry, were used to obtain high contrast optically bistable switches in various schemes utilizing both "normally on" and "normally off" types of modulators. The flexibility provided by the inverted structure in realizing novel optical and electrical interconnections is demonstrated.

## INTRODUCTION

Symmetric Fabry-Pérot (ASFP) light modulators are of interest for their high contrast ratio and large throughput. The growth of InGaAs/GaAs based ASFP modulators has been especially challenging since they require not only the high reflectivity Bragg mirrors and very accurate ( $<1\%$ ) thickness control, but also growth of high quality strained multiple quantum wells (MQW). Based on our previous success in the growth of such InGaAs/GaAs MQW reflection high-energy electron diffraction (RHEED) determined growth conditions,<sup>1,2</sup> we have realized InGaAs based ASFP modulators with contrast ratio as high as 66:1 and 30% attendant throughput.<sup>3,4</sup> The advantages of InGaAs based ASFP modulators on GaAs substrates have been particularly afforded in the inverted geometry<sup>5,6</sup> in which light passes in and out through the substrate. Such a geometry allows replacing the growth of the high reflectivity Bragg mirror ( $\sim 2.5 \mu\text{m}$  thick) by an externally deposited mirror as well as postgrowth tuning of the Fabry-Pérot cavity. This geometry also allows hybrid integration with Si circuitry through flip-chip bonding. In addition, in efforts towards all III-V integration, such a geometry provides simultaneous optical access from both sides of the modulator/detector chip.

The objective of this article is to explore realization of high contrast bistable switches using both "normally on" and "normally off" ASFP modulators, especially in the inverted geometry. For "normally on" modulators, bistable switching has been demonstrated in the monolithic optoelectronic transistor (MOET) configuration proposed<sup>7</sup> by Williamson and co-workers. In this configuration [Fig. 1(a)], a tunneling diode and a field effect transistor (FET) are used with two pixels of the modulator/detector chip to obtain bistability and gain. For "normally off" modulators, bistable switching has been achieved in various self-electrooptic device (SEED) configurations.<sup>8</sup> In the usual SEEDs, the negative differential resistance (NDR) required for bistable operation is based on physical effects that cause reduction in the absorption coefficient at the operating wavelength due to an excitonic red shift under an applied electric field. By contrast, in the present work the NDR is obtained from the blue shift of the Fabry-Pérot mode due

to the reduction of refractive index near the exciton shoulder. Additionally, we show that it is possible to integrate a binary phase Fresnel microlens array with the inverted spatial light modulators (SLM) for lensless free space optical interconnection.

## II. EXPERIMENTAL

The samples were grown in the USC RIBER 32P solid source molecular-beam epitaxy (MBE) system under RHEED determined growth conditions.<sup>2</sup> The conventional ASFP modulator consists of a 17.5 period  $n^+$  GaAs(654 Å)/AlAs(784 Å) Bragg Mirror, a 50 period In<sub>0.11</sub>Ga<sub>0.89</sub>As(100 Å)/GaAs(125 Å) MQW, and a 5 period  $p^+$  Bragg mirror grown on a  $n^+$  GaAs substrate. The inverted modulator consists of a 5.5 period  $n^+$  Bragg mirror, the same MQW as above, and an externally deposited Au mirror, grown on a semi-insulating GaAs substrate. The growth temperatures for the first mirror and the MQW were 600 and 545 °C, respectively, in both the structures and 555 °C for the second mirror in the conventional structure. The growth rates were 0.76 and 0.60 monolayer (ML)/s for the GaAs and AlAs in the mirror, and 0.26 and 0.29 ML/s for the GaAs and InGaAs in the MQW. The resonant tunneling diode (RTD) consists of a triple-well, double-barrier structure with two 16 ML In<sub>0.25</sub>Ga<sub>0.75</sub>As spacers, two 10 ML AlAs barriers, and one 18 ML In<sub>0.25</sub>Ga<sub>0.75</sub>As well.<sup>9</sup>

## III. RESULTS

The modulators/detectors used in the MOET switch [Fig. 1(a)] are two pixels of the InGaAs/GaAs conventional ASFP modulator array. The modulator pixel produces a contrast ratio of 20:1, and a reflectivity change of 44.4% at 14 V reverse bias when the incident light power is below 0.4 mW. When the incident power is increased, a higher voltage is required to compensate for the voltage drop across the contact due to increased photocurrent. The InGaAs/AlAs RTD of Sec. II and a commercial Si-FET were used. The 12 μm × 12 μm cross-section RTD shows a peak current  $I_p = 0.6 \text{ mA}$  and a valley current of  $I_v = 0.2 \text{ mA}$ . Details of the RTD and the modulator characteristics are reported in Refs. 9 and 10, respectively. Figure 1(b)

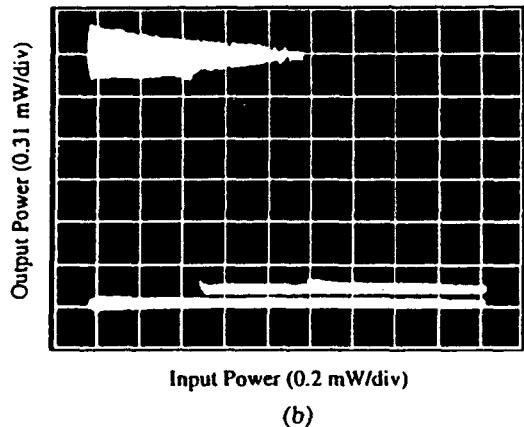
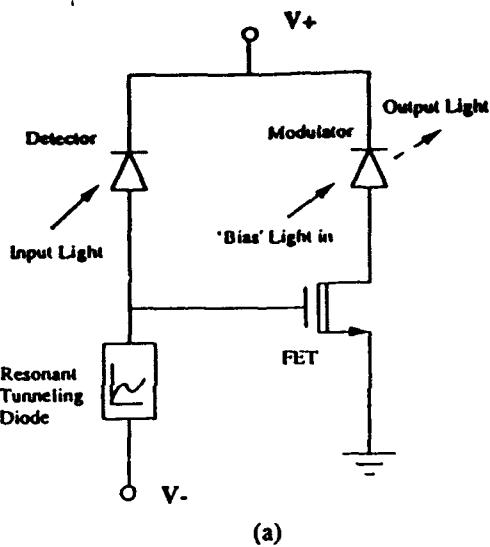


FIG. 1. (a) Circuit diagram of the monolithic optoelectronic transistor configuration (after Ref. 4). (b) Optical switching performance of the MOET for supply a voltage  $V_+ = 20$  V and  $V_- = -1$  V. The lowest trace corresponds to the zero output intensity level.

shows the input/output characteristics of this switch. At  $V_+ = 20$  V and  $V_- = -1$  V, a contrast ratio of 20:1 and a  $\Delta R$  of 44.4% are obtained, the same as the modulator pixel employed. A fan-out of 2 is realized. The fan-out can be significantly increased by reducing the peak current of the RTD. Recently, we have successfully decreased the peak current by an order of magnitude by reducing the RTD cross section to  $5 \mu\text{m} \times 5 \mu\text{m}$  and increasing the RTD barrier thickness. Moving towards an all-(III-V) MOET, we have also examined the use of a GaAs metal-semiconductor field effect transistor (MESFET) in place of the Si-FET. A contrast ratio of 4:1 is realized at the 12 V maximum bias that the MESFET could sustain. Further improvement requires redesign of either the GaAs FET for higher voltage operation or the modulator for lower voltage operation.

Turning to SEEDs, "normally off" pixels of modulator of Sec. II were employed in conjunction with a Si photodiode (D-SEED), Si phototransistor (T-SEED), or another modulator pixel (S-SEED). The pixel (10,10) has a

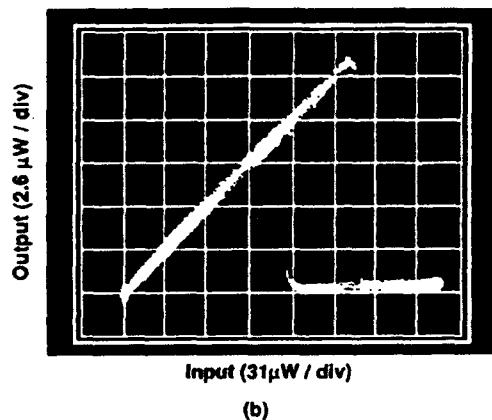
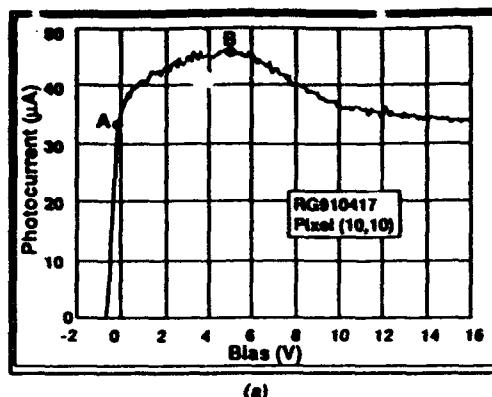


FIG. 2. (a) Photocurrent behavior of pixel (10, 10) as a function of the applied reverse bias with  $150 \mu\text{W}$  of illumination at  $9484 \text{ \AA}$ . (b) The D-SEED switching characteristics of pixel (10, 10). The light beam to the modulator was scanned.

FP mode at  $9484 \text{ \AA}$  with a reflectivity of  $\sim 0.4\%$  at zero bias. At 20 V reverse bias, the FP mode blue shifts to  $9451 \text{ \AA}$  and the reflectivity at  $9484 \text{ \AA}$  increases to  $\sim 20\%$ . This is manifest in the photocurrent  $I$ - $V$  characteristics obtained under  $9484 \text{ \AA}$  illumination [Fig. 2(a)], which shows onset of NDR at 5 V. Figure 2(b) shows the switching input/output characteristics of the pixel using a Si photodiode as the load. A contrast ratio of  $\sim 50:1$  and a  $\Delta R$  of  $\sim 20\%$  is obtained for this D-SEED.

To compare the characteristics of the D-SEED and the T-SEED, Figs. 3(a) and 3(b) show the photo  $I$ - $V$  characteristics of the Si photodiode and the Si phototransistor used in conjunction with modulator pixel (11,11). To obtain the same current output from the photodiode and the phototransistor, the light power to the phototransistor was reduced by a factor of 200. Note that the photodiode has a larger differential resistance than the phototransistor under reverse bias. This difference is reflected in the switching input/output characteristics of the D-SEED and the T-SEED, shown in Figs. 3(c) and 3(d), respectively. The D-SEED has a narrower hysteresis but well maintained on and off states. The hysteresis in the T-SEED is much wider and has a well maintained logic state only in its central region. The contrast ratios of the D-SEED and the T-SEED are nearly the same. Though not shown here, two modulator pixels were used to construct the first S-SEED

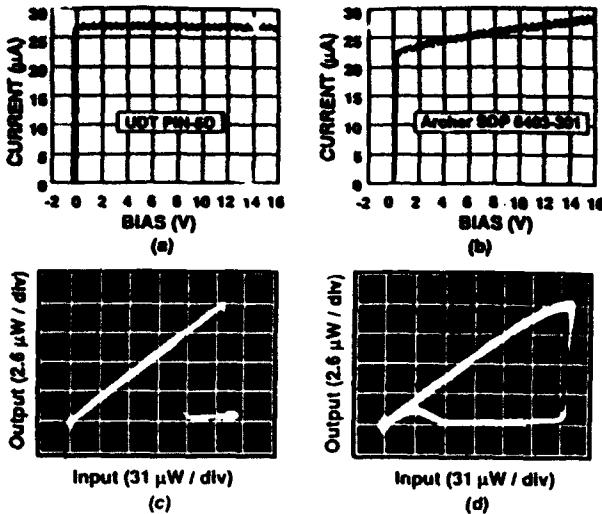


FIG. 3. Panels (a) and (b) show the photocurrent  $I$ - $V$  characteristics of a UDT PIN-5D Si photodiode and Archer SDP 8403-301 Si phototransistor. Note that in order to obtain the same current output from the photodiode and the phototransistor, the light power to the phototransistor was reduced by a factor of 200. Panels (c) and (d) show the corresponding D-SEED and T-SEED switching characteristics for pixel (11, 11). The beam to the modulator was scanned.

based on the inverted ASFP modulator. A bistable switching contrast ratio of 5:1 was achieved, comparable to that reported for the unstrained GaAs/AlGaAs MQW based S-SEEDs. With improved device processing this figure can be improved to as good as that of the D-SEED. Recently direct growth of a heterojunction phototransistor on the inverted InGaAs ASFP modulator to achieve a monolithically integrated structure has also been successfully realized.<sup>11</sup>

Finally, we turn to the fabrication of binary phase Fresnel microlenses for integration with the modulator array in the *inverted* geometry. Figure 4(a) shows an array of nine sets of binary phase Fresnel lenses fabricated using focused ion beam. These lenses are a simplification of Fresnel lens and consist of alternating concentric zones in two discrete plateaus rather than a continuous manner.<sup>12</sup> The center to center distance between lenses in the picture is 37  $\mu\text{m}$  in both directions. An array of focused spots was clearly seen on the first order focal plane [Fig. 4(b)]. Further studies regarding the design of the lens, the transfer efficiency, the intensity profile at the focal plane, stray light level, and cross talk between the lenses are in progress.<sup>13</sup>

In conclusion, high contrast ratio optically bistable switches are demonstrated using strained InGaAs based

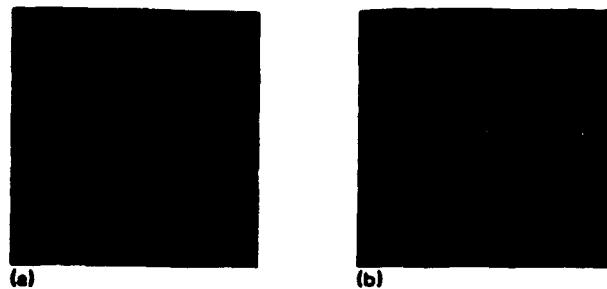


FIG. 4. (a) Nomarski micrograph of an array of nine sets of binary phase Fresnel lenses made via focused ion beam. (b) Photograph of the incident light taken at the focal plane of the lenses.

ASFP modulators in both "normal" and "inverted" geometry and for both "normally on" and "normally off" types of modulators. The flexibility of the inverted modulator structure is explored and the possibility of integrating a microlens array and SLM is demonstrated.

#### ACKNOWLEDGMENTS

This work was supported by the AFOSR under the URI Center for Optical Computing, by DARPA, by the JSEP program, and by the NCIPT (DARPA).

- <sup>1</sup>Li Chen, K. C. Rajkumar, and A. Madhukar, *Appl. Phys. Lett.* **57**, 2478 (1990).
- <sup>2</sup>P. Chen, J. Y. Kim, A. Madhukar, and N. M. Cho, *J. Vac. Sci. Technol. B* **4**, 890 (1986).
- <sup>3</sup>Kezhong Hu, Li Chen, A. Madhukar, P. Chen, K. C. Rajkumar, K. Kaviani, Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., *Appl. Phys. Lett.* **59**, 1108 (1991).
- <sup>4</sup>Li Chen, Kezhong Hu, K. C. Rajkumar, S. Guha, R. Kapre, and A. Madhukar, in *Materials Research Society Symposium Proceedings*, edited by C. Warde, J. Stamatoff, W. Wang (Materials Research Society, Pittsburgh, PA, 1991), Vol. 228, p. 213.
- <sup>5</sup>Kezhong Hu, Li Chen, A. Madhukar, P. Chen, C. Kyriakakis, Z. Karim, and A. R. Tanguay, Jr., *Appl. Phys. Lett.* **59**, 1664 (1991).
- <sup>6</sup>Kezhong Hu, Li Chen, R. Kapre, K. C. Rajkumar, A. Madhukar, Z. Karim, C. Kyriakakis, A. R. Tanguay, Jr., in *Proceedings of the IEEE-LEOS 1991 Summer Topical Meeting* (IEEE, New York, NY, 1991), p. 18.
- <sup>7</sup>R. C. Williamson, in *Conference on Lasers and Electro-optics* (Optical Society of America, Washington, DC, 1989), p. 32.
- <sup>8</sup>A. L. Lentine, H. S. Hinton, D. A. B. Miller, J. E. Henry, J. E. Cunningham, and L. M. F. Chirovsky, *IEEE J. Quantum Electron.* **25**, 1928 (1989) and references therein.
- <sup>9</sup>R. M. Kapre, Kezhong Hu, Li Chen, S. Guha, and A. Madhukar, in *Materials Research Society Symposium Proceedings*, edited by C. Warde, J. Stamatoff, and W. Wang (Materials Research Society, Pittsburgh, PA, 1991), Vol. 228, p. 219.
- <sup>10</sup>Li Chen, R. M. Kapre, Kezhong Hu, and A. Madhukar, *Appl. Phys. Lett.* **59**, 1523 (1991).
- <sup>11</sup>Kezhong Hu, Li Chen, and A. Madhukar, *IEEE Photon. Technol. Lett.* (in press).
- <sup>12</sup>K. Rastani, A. Marrakchi, S. F. Habiby, W. M. Hubbard, H. Gilchrist, and R. E. Nahory, *Appl. Opt.* **30**, 1347 (1991).
- <sup>13</sup>Wei Chen, Li Chen, Ping Chen, and A. Madhukar (unpublished).

# **Effects in strained $In_{0.2}Ga_{0.8}As/GaAs$ multiple quantum wells on patterned and unpatterned substrates: A near-infrared cathodoluminescence study**

D. H. Rich, K. C. Rajkumar, Li Chen,<sup>a)</sup> and A. Marinkar<sup>b)</sup>

Photonic Materials and Devices Laboratory, Department of Materials Science and Engineering,  
University of Southern California, Los Angeles, California 90089-0241

T. George, J. Maserjian, and F. J. Grunthaner

Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute  
of Technology, Pasadena, California 91109

A. Larsson

Chalmers University of Technology, Department of Optoelectronics and Electrical Measurements,  
S-412 96 Göteborg, Sweden

(Received 6 February 1992; accepted 20 March 1992)

The spatial distribution of the long-wavelength luminescence in thick  $In_{0.2}Ga_{0.8}As/GaAs$  multiple quantum wells (MQWs) grown on patterned and unpatterned substrates has been investigated using cathodoluminescence (CL) imaging and spectroscopy. By spatially correlating the luminescence arising from the MQW exciton recombination ( $\lambda \approx 950$  nm) with the longer wavelength ( $1000 \leq \lambda \leq 1200$  nm) luminescence arising from the defect-induced recombination, we demonstrate that it is possible to determine the regions of highest film quality in both the mesa and valley regions for growth on patterned GaAs substrates. The present approach enables a judicious determination of the optimal regions to be used for active pixels in InGaAs/GaAs devices. For growth on unpatterned substrates, the CL spectra show defect-induced broad bands between  $1000 \leq \lambda \leq 1600$  nm. These bands exhibit spatial variations which correlate with the dark line defects (DLDs) observed in the  $\lambda = 950$  nm exciton luminescence imaging. Transmission electron microscopy showed that [110]-oriented misfit dislocations occur primarily at the substrate-to-MQW interface. The large spatial variation of the luminescence intensities indicates that the DLDs observed in CL images are caused by the presence of nonradiative recombination centers occurring in the MQW region located above the interface dislocations. This study provides new information describing the origin and nature of DLDs and differs from previous models, which have regarded the electronic nature of dislocation cores as the primary mechanism for inducing DLD radiative contrast in luminescence imaging of strained InGaAs/GaAs.

## **INTRODUCTION**

Sently, there is a considerable interest in using selective epitaxial growth approaches to achieve the fabrication of thick coherent films of strained InGaAs and related multiple quantum well structures (MQW) structures on As substrates. Many applications in the area of optical computing, communication, and detection will benefit from the continuing advancements made in the molecular-beam epitaxial (MBE) growth techniques of highly strained InGaAs films. Previous studies employing transmission electron microscopy (TEM),<sup>1-4</sup> photoluminescence (PL),<sup>2,5</sup> cathodoluminescence (CL),<sup>6</sup> optical absorption,<sup>1,3</sup> and micro-Raman,<sup>4</sup> have demonstrated a substantial reduction in the density of misfit dislocations occurring in thick  $In_xGa_{1-x}As$  films and MQWs grown on As mesas with dimensions between 1 and  $100\text{ }\mu\text{m}$ . The technique of CL microscopy enables the measurement of luminescence behavior with a scale of approximately micron or less; this is quite valuable for the optical characterization of III-V semiconductor device arrays having active pixels with such dimensions. The spatial distribution of the near-infrared ( $900 \leq \lambda \leq 1600$  nm) luminescence in thick  $In_{0.2}Ga_{0.8}As/GaAs$  MQWs grown on patterned and unpatterned substrates has been examined with CL imaging and spectroscopy. We show that by spatially correlating the luminescence arising from electron-to-heavy hole (e-hh) exciton recombination in the MQWs with the longer wavelength luminescence arising from the defect-induced recombination, it is possible to determine the regions of highest film quality in both the mesa and valley regions for patterned InGaAs/GaAs MQW structures. The effects of cation migration near faceted sidewalls of chemically etched rectangular mesas are examined.

For growth of MQWs on unpatterned substrates, the long-wavelength luminescence features are found to correlate spatially with the dark line defects (DLDs) seen in the imaging of the MQW exciton recombination at  $\lambda \approx 950$  nm. Plan-view and cross-sectional TEM show that misfit dislocations in the samples are confined to the region near the interface. The luminescence intensities from various MQW structures are quantified, and a large variation in intensity of the luminescence associated with the DLDs suggest that most of the QWs in the vicinity of the DLDs are affected by the misfit dislocations. This intensity vari-

ation is found to be relatively independent of the electron beam probing depth, and indicates that the QWs are affected homogeneously throughout all the layers, in contrast to the presence of misfit dislocations which occur only at the MQW-to-substrate interface. This luminescence behavior is attributed to a relative increase in the number of nonradiative recombination centers occurring in regions above the interface dislocations. These results suggest that nonradiative point defects above the interface dislocations are left in the wake of dislocation propagation and multiplication. In the Si system Kimerling and co-workers,<sup>7</sup> using deep-level transient spectroscopy, have speculated that it is the point defects gettered around dislocations which are largely responsible for defect induced midgap states and the associated recombination properties. Similarly, we demonstrate that a Cottrell atmosphere of point defects surrounding dislocations is the most reasonable explanation for the DLD behavior observed in this study of In<sub>0.2</sub>Ga<sub>0.8</sub>As/GaAs MQW structures.

## II. EXPERIMENT

MQW samples were grown by MBE using standard In, Ga, and As sources. Our study of growth on unpatterned substrates involves 65 Å thick In<sub>0.2</sub>Ga<sub>0.8</sub>As QWs surrounded by GaAs barriers. In samples designated D92, D18, and D38, 44 periods of In<sub>0.2</sub>Ga<sub>0.8</sub>As MQWs were grown with barrier thicknesses of 780, 400, and 115 Å, respectively. In sample D179, a 65 Å MQW structure having 100 and 1230 Å barriers with 14 periods (28 QWs) were grown. A control sample, D155, was grown with one 65 Å thick QW. One of the samples, D92, was periodically  $\delta$  doped; Be and Si doping planes with concentrations of  $9.0 \times 10^{12}$  and  $3.0 \times 10^{12} \text{ cm}^{-2}$  were inserted at the centers and 290 Å off centers of the GaAs barriers to produce a *nipi* effect. This sample was grown originally for the purpose of studying suitable structures for optically addressed spatial light modulators.<sup>8</sup> The spatial separation of electrons and holes enables a large optically induced absorption modulation for this structure; this has been recently reported.<sup>8,9</sup> For patterned growth, a  $n^-$ -type GaAs(100) substrate was patterned using conventional optical lithography and wet chemical etching.<sup>1-3</sup> The crystallographically selective etch results in undercut and inclined side-wall planes [mostly {111} with some high index facets] at the mesa edges. The resulting mesas had dimensions of about 2 μm thick × 16 μm × 18 μm with a 40 μm spacing in both directions before MBE growth. A scanning electron micrograph (SEM) image of a typical mesa is shown in Fig. 1. About half the wafer was left unpatterned in order to allow for examination of a reference sample. The structure grown on top consisted of a 100 period In<sub>0.2</sub>Ga<sub>0.8</sub>As (80 Å)/GaAs (160 Å) MQW (having ~2.38 μm total thickness) with a 5000 Å  $p^+$ -type GaAs capping layer. Cross-sectional TEM (XTEM) and absorption results of this sample (designated RG891110) have been previously described.<sup>1,3</sup> CL measurements were performed with a JEOL 840-F field-emission SEM at the Jet Propulsion Laboratory.<sup>10</sup> A North Coast E0-817L Ge  $n-i-p$  detector was used to measure the signal dispersed by a 0.25 m

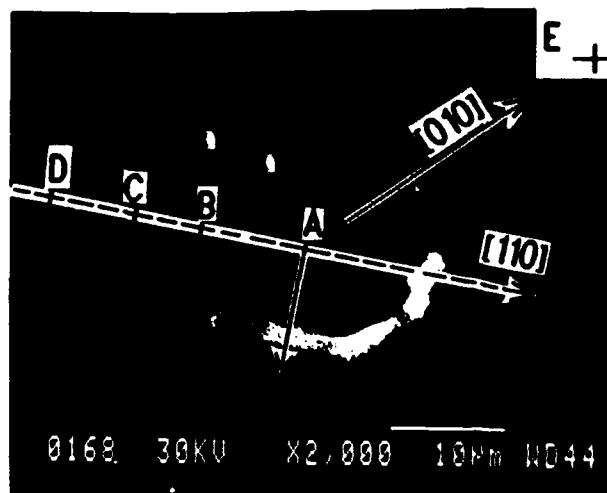


FIG. 1. SEM of a typical mesa of the In<sub>0.2</sub>Ga<sub>0.8</sub>As/GaAs MQW structure. The dashed and solid lines with arrows oriented along [110] and [1-10], respectively, indicate the position of the electron beam during line scans. Points A-E indicate beam position during local CL spectroscopy.

monochromator. An electron beam current of 0.5 nA at accelerating voltages ranging from 10 to 40 kV was used to probe the sample for the CL measurements. The sample was maintained at a temperature of ~77 K.

## III. RESULTS AND DISCUSSION

### A. In<sub>0.2</sub>Ga<sub>0.8</sub>As growth on patterned GaAs(100)

A SEM image of the structure showing a 220 μm × 170 μm area is presented in Fig. 2(a). Scanning monochromatic CL images corresponding to the same region are shown in Figs. 2(b)-2(d) for wavelengths of 950, 1040, and 1120 nm, respectively. Regions of increasing luminescence signal are represented by areas of lighter shades. The image with  $\lambda = 950 \text{ nm}$  [in Fig. 2(b)] shows a hatched region of bright luminescence with maximum intensity in the valley regions midway along the [010] diagonal (dotted line in Fig. 1) between mesa centers. There is some observable emission from the centers and edges of the mesas at this wavelength. A dark halo region is found to surround the mesas, where little emission is detected. As the wavelength is increased to 1040 and 1120 nm, the images in Fig. 2 show an enhanced emission from the centers and sides of the mesas relative to the valley regions. Also, for these wavelengths (and for other CL imaging wavelengths in the region  $1000 < \lambda < 1200 \text{ nm}$  not shown here), the luminescence patterns on top of most of the mesas show a distinct dumbbell shape with maximal intensity near the mesa edges parallel to [110], as seen in Figs. 2(c) and 2(d). The dumbbell shape on the mesas, as seen in the  $\lambda > 1000 \text{ nm}$  CL images and the [110] CL line scan, indicates the presence of interfacet cation migration in which the In migrates up along the inclined edges. The presence of contiguous low index planes is known to cause interfacet migration; this behavior has been observed to cause compositional variations in the AlGaAs and InGaAs films grown on patterned substrates.<sup>1-3,8,11</sup>

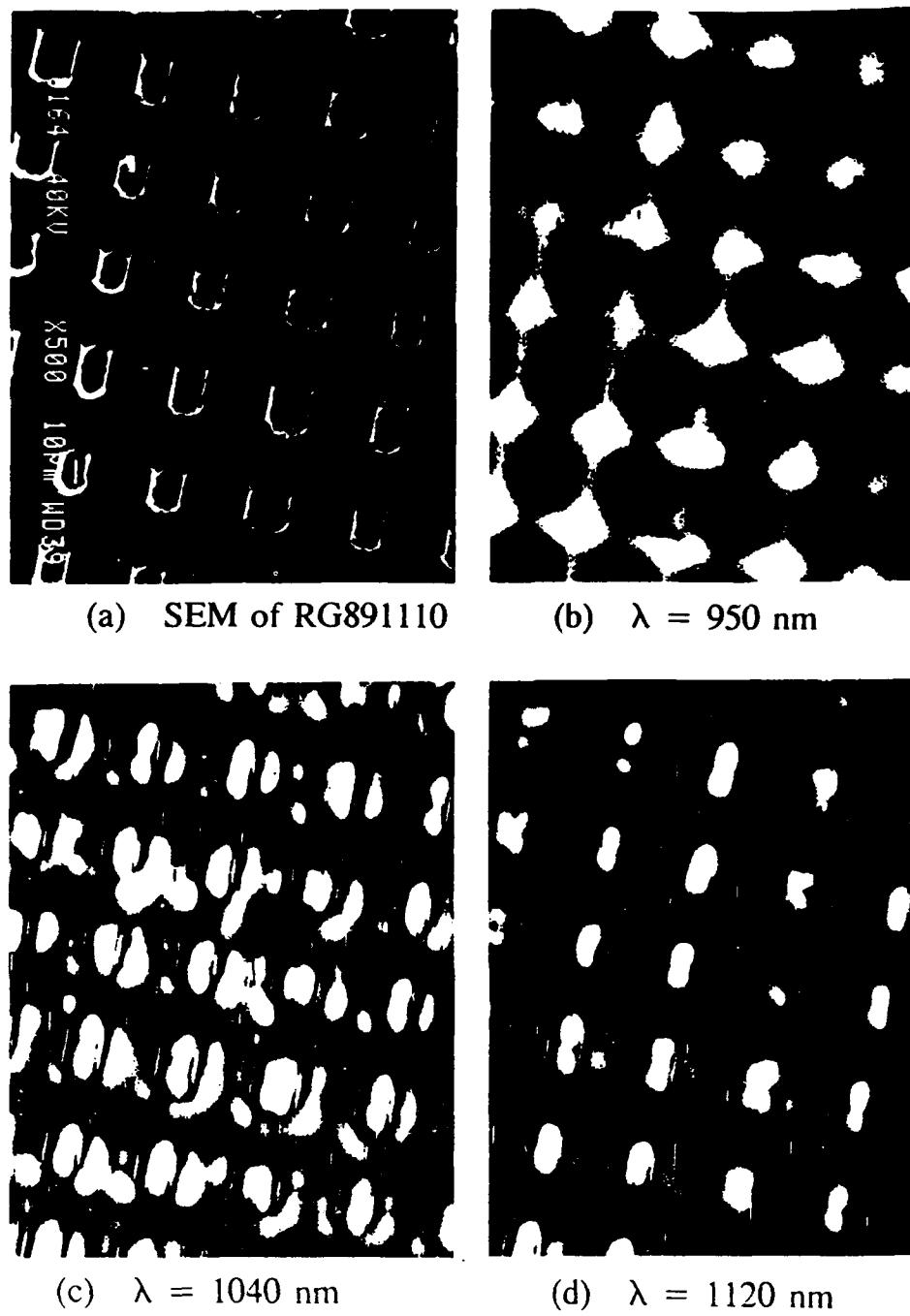


FIG. 2. SEM (a) and scanning monochromatized cathodoluminescence images [(b), (c), and (d) correspond to wavelengths of 950, 1040, and 1120 nm, respectively] of the same region of the  $\text{In}_{0.2}\text{Ga}_{0.8}\text{As}/\text{GaAs}$  MQW structure.

Previously, TEM measurements of the mesa region for RG891110 has shown that the linear defect density to be less than  $\sim 10^5 \text{ cm}^{-1}$ . Since the maximum spatial resolution in CL is limited by the carrier diffusion length ( $\sim 1 \mu\text{m}$  for GaAs), we do not expect to see the effects of individual dislocations. It will be shown below (in Sec. III B) that for the D-series unpatterned samples, the presence of underlying misfit dislocations at the MQW-to-substrate interface gives rise to a crosshatched pattern when the linear defect density is less than  $\sim 10^5 \text{ cm}^{-1}$ .

Localized CL spectra are shown in Fig. 3; locations

A-E correspond to points on a mesa center, mesa edge, valley region near mesa, midway between mesas along [110], and midway between mesas along [010], respectively, as labeled in Fig. 1. The spectrum labeled Ref. corresponds to an unpatterned region and does not exhibit a measurable exciton luminescence. The peak at  $\lambda = 950 \text{ nm}$  corresponds to the e-hh exciton luminescence. The regions exhibiting the largest defect-induced luminescence ( $1000 \leq \lambda \leq 1200 \text{ nm}$ ) on the mesa are found in close proximity to the inclined regions at the mesa edges (position B in Fig. 1) which result from the crystallographic selectivity

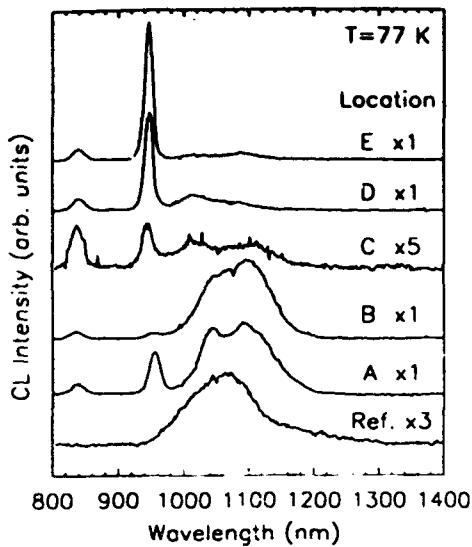


FIG. 3. Localized CL spectroscopy at the points A-E shown in Fig. 1. The sample labeled Ref. corresponds to the spectrum taken in the non-patterned region of the In<sub>0.2</sub>Ga<sub>0.8</sub>As/GaAs MQW structure.

of the etch. The region possessing the highest film quality has been determined by measuring the exciton-to-defect emission ratio for a variety of positions in the structure. The data show, surprisingly, that the highest-quality film exists in the valley regions which are midway between the mesa centers (location E in Fig. 1) and not at the mesa centers (location A in Fig. 1). These results are expected to influence current thinking with regard to the design and processing of pixelated devices.

#### B. Growth of In<sub>0.2</sub>Ga<sub>0.8</sub>As on unpatterned GaAs(100)

TEM was performed on the D38, D18, D92, and D179 samples. The MQW regions showed well-ordered In<sub>0.2</sub>Ga<sub>0.8</sub>As and GaAs which were visibly free of structural defects. Only at and below the MQW-to-substrate interface did there appear to be disorder in the form of interface misfit dislocations and looping dislocations that propagated into the GaAs substrates. Furthermore, by progressively thinning away the back side of the substrates, we determined, using plan view TEM, that the misfit dislocation cores were confined solely to the MQW/GaAs interfacial region.

CL images of D38, D18, D92, D155, and D179 are shown in Figs. 4(a)-4(e). The wavelength used for the imaging was  $\lambda \approx 950$  nm and corresponds to the e-hh transition in the MQW samples. The DLDs are evident in each of the images except for D155, the single QW sample [Fig. 4(d)]. The absence of misfit dislocations in the single QW sample is to be expected since the nominal critical thickness of a single In<sub>0.2</sub>Ga<sub>0.8</sub>As layer grown on GaAs(100) is  $\sim 150\text{--}200\text{ \AA}$ .<sup>12</sup> These images showing DLDs here are similar to luminescence images of thick InGaAs films previously studied.<sup>6</sup>

In order to further assess the luminescence properties, we examined CL spectra of the unpatterned InGaAs sam-

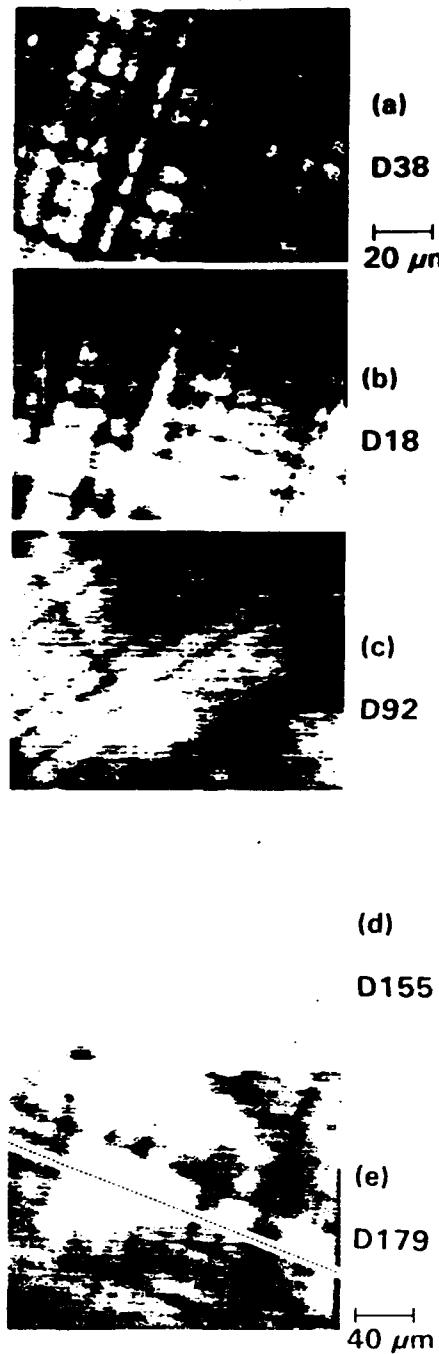


FIG. 4. Scanning monochromatic CL images of the e-hh excitor luminescence for  $\lambda \approx 950$  nm. The sample number and MQW structure are described in (a)-(e). The 20  $\mu\text{m}$  scale indicated in (a) also represents the scale for (b)-(d); note the factor of 2 scale change in (e). A dashed line along [110] in (e) indicates the electron beam path during the line scan measurements for D179.

ples in the wavelength range  $900 \leq \lambda \leq 1600$  nm. The spectra (not shown here) are similar to that shown in Fig. 3 for the growth on the patterned substrates; four distinct features in the luminescence were observed centered at 950, 1250, 1380, and 1460 nm. The peak at  $\lambda = 950$  nm is the e-hh transition. Joyce *et al.*<sup>13</sup> using PL, have observed emission in the  $1000 \leq \lambda \leq 1600$  nm range for thick InGaAs films grown on GaAs. These emissions were found to in-

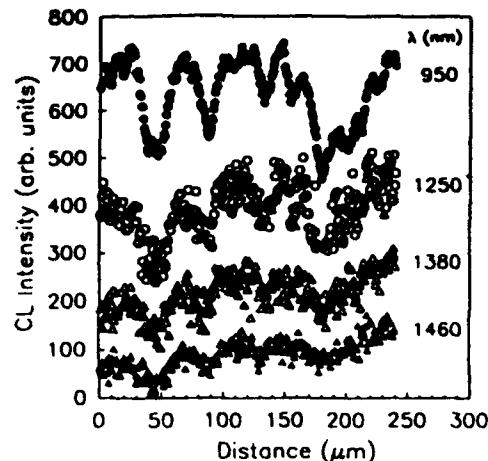


FIG. 5. CL line scans (intensity vs position histogram) for MQW sample D179. The electron beam position is indicated in Fig. 4. The peaks and valleys in the  $\lambda = 950$  nm scan of the e-hh exciton luminescence correlates with the peaks and valleys of the longer wavelength 1250, 1380, and 1460 nm scans.

crease markedly when the In<sub>0.17</sub>Ga<sub>0.83</sub>As layer thicknesses exceeded the critical thickness, and Joyce *et al.* concluded that this behavior was due to an increase in interface defects, such as misfit dislocations, beyond the critical thickness.<sup>13</sup> In order to test this hypothesis, we have determined the spatial distribution of all distinct emissions in the wavelength range  $900 \leq \lambda \leq 1600$  nm. We have measured intensity versus position histograms or line scans (see Fig. 5) which quantitatively give the spatial variation in luminescence intensity for the different wavelengths. The line scans for wavelengths of 1250, 1380, and 1460 nm have essentially the same spatial variation as the  $\lambda = 950$  nm scan, i.e., peaks and valleys in the longer wavelength scans correspond with peaks and valleys in the  $\lambda = 950$  nm scan. This is a surprising result since, from Ref. 12, we would expect the region around the DLDs to yield an enhanced emission rather than the reduced emission at longer wavelengths.

The nature of the dislocation cores (i.e., electronic structure and reconstruction) in strained III-V materials is known to effect the recombination process. The studies of Fitzgerald *et al.*<sup>6</sup> and Petroff *et al.*<sup>14</sup> on strained films have shown that 60° dislocations and edge dislocations have different nonradiative recombination rates in the vicinity of these dislocations. It is apparent from the data of Fig. 5 that the nonradiative mechanisms responsible for the reduction in the e-hh exciton luminescence are also responsible for reduction in the  $1000 \leq \lambda \leq 1600$  nm emissions. This results in longer wavelength DLDs spatially correlated with the conventional e-hh exciton DLDs. The origin of the  $1000 \leq \lambda \leq 1600$  nm emissions is, evidently, not directly caused by the presence of misfit dislocations and may be due to the presence of point defects. Similar broad and midgap emissions have been observed in thick and heavily strained GaAs films grown on Si which are believed to contain interstitial, antisite, and vacancy defects.<sup>15</sup> In addition, we can not rule out emissions due to the presence of

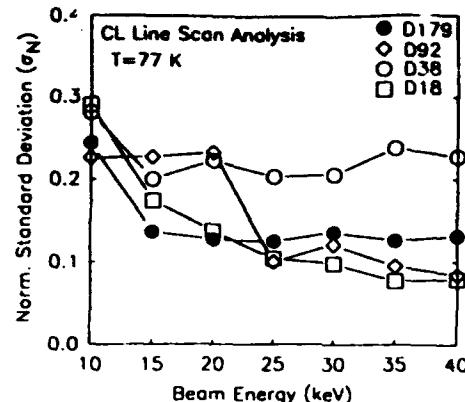


FIG. 6. The normalized standard deviation ( $\sigma_N$ ) of the e-hh exciton luminescence line scans for MQW samples. The standard deviation has been normalized to the mean of the exciton luminescence for each line scan. The large magnitude of the variation (see discussion) precludes the possibility of only one QW in each structure being affected by a single dislocation core at the MQW-to-substrate interface.

In<sub>0.2</sub>Ga<sub>0.8</sub>As/GaAs interfaces (i.e., recombinations involving interface states). Other types of structural defects can also induce midgap emissions. A particularly striking result in Fig. 5 is the large intensity variation seen in the line scan data. The variation of the luminescence in the  $\lambda = 950$  nm line scans can be quantified by evaluating the standard deviation of all points shown in a line scan, and normalizing the standard deviation with respect to the mean. In order to reduce the uncertainties due to variation in the signal-to-noise caused by changes in the luminescence signal associated with changing the excitation conditions (current and beam energy) and different sample thicknesses, a cubic spline fit was used to determine the "smooth curve" through each line scan. The normalized standard deviation  $\sigma_N$  of fitted  $\lambda = 950$  nm line scans are shown in Fig. 6 for all samples with a beam energy ranging from 10 to 40 keV. A general trend is observed in which  $\sigma_N$  is fairly constant over a large energy range and increases slightly with a decreasing probing depth for energies less than about 20 keV. The total thicknesses of the MQW region of D179, D92, D38, and D18 are 2.0, 3.7, 0.8, and 2.0  $\mu\text{m}$ , respectively, and a given variation in the probing-depth will affect different regions of each sample. From the above TEM results, we know that misfit dislocations occur only in the region at the MQW-to-substrate interface. We should expect, therefore, that only recombination in the QW closest to the interface will be disturbed directly by the presence of a misfit dislocation. A predicted  $\sigma_N$  for a 44 MQW structure in which the luminescence from one QW at the interface is completely suppressed by a misfit dislocation, assuming a homogenous carrier excitation and a simple sinusoidal spatial variation in the luminescence, is  $\sigma_N = 0.008$ . This is at least an order of magnitude too low to explain the present results of Fig. 6. Instead the magnitude of  $\sigma_N$  and its relative constancy for each sample in Fig. 6 suggest that the nonradiative mechanisms giving rise to DLDs are spread homogeneously throughout all MQWs in each sample, and the effects become slightly more pro-

anced for more surface sensitive probing conditions. These results have significant ramifications concerning the origin of the DLDs seen in this and previous studies of thick InGaAs films grown on GaAs. In previous studies, it was simply assumed that the electronic nature of the misfit dislocations were responsible for the creation of competing nonradiative channels.<sup>6,14</sup> In that model, nonradiative minority carrier recombination occurs at the reconstructed dislocation cores as a result of localized band bending in the vicinity of the cores; the density of defect-induced states in the band gap and doping concentration will determine local depletion region sizes. However, the present data suggest that carrier recombination at the dislocation cores is a minor effect in producing the DLDs observed in C. This can be proven immediately by considering the *nipi* structure of sample D92. The modulation doping will cause electrons and holes to spatially separate for carriers which do not recombine at the location of the In<sub>0.2</sub>Ga<sub>0.8</sub>As QWs, and this results in negligible carrier diffusion along the [100] growth direction.<sup>8,9</sup> Therefore, it is apparent that recombination of carriers which are generated more than one MQW period away from the dislocation core would be negligible in this model. The only reasonable explanation for the behavior of  $\sigma_N$  is as follows. The existence of nonradiative recombination centers spread homogeneously in the MQW material above the dislocations is the primary cause for the large variation in the MQW exciton luminescence. The nonradiative centers will evidently compete with the longer wavelength 1000  $\lesssim \lambda \lesssim$  1600 nm radiative channels, resulting in the spatial correlation of DLDs imaged at  $\lambda \approx 950$  nm with DLDs imaged at longer wavelengths (as illustrated in Fig. 5 for line scans). Dislocation propagation and multiplication such as, e.g., the Hagen-Strunk<sup>16</sup> and surface half-loop nucleation<sup>17</sup> mechanisms are expected to disturb the material located above the dislocations. Heavily dislocated materials, such as that seen in the strained GaAs/Si system are known to contain a significant density of point defects.<sup>7,15</sup> It is, therefore, likely that a "Cottrell atmosphere" of point defects left in the wake of dislocation propagation and multiplication coupled with the concomitant epitaxial growth over the vicinities of such propagation are responsible for the observed behavior of  $\sigma_N$ .

## ACKNOWLEDGMENTS

The research described in this article was performed by the Center for Space Microelectronics Technology, Jet

Propulsion Laboratory, California Institute of Technology, and was jointly sponsored by the Defense Advanced Research Projects Agency, the Strategic Defense Initiative Organization, Innovative Science and Technology Office, and the National Aeronautics and Space Administration, Office of Aeronautics, Exploration, and Technology. At USC, part of this work was supported by AFOSR, URI (AFOSR), DARPA, and ONR.

<sup>1</sup>Also with the Department of Physics.

<sup>2</sup>A. Madhukar, K. C. Rajkumar, L. Chen, S. Guha, K. Kaviani, and R. Kapre, *Appl. Phys. Lett.* **57**, 2007 (1990).

<sup>3</sup>S. Guha, A. Madhukar, and L. Chen, *Appl. Phys. Lett.* **56**, 2304 (1990).

<sup>4</sup>L. Chen, K. Hu, K. C. Rajkumar, S. Guha, R. Kapre, and A. Madhukar, *Proceeding of the Materials Research Society Symposium on Materials for Optical Information Processing*, Anaheim, CA, May 1-3, 1991 (in press).

<sup>5</sup>W. C. Tang, H. J. Rosen, S. Guha, and A. Madhukar, *Appl. Phys. Lett.* **58**, 1644 (1991).

<sup>6</sup>Y. Zou, P. Grodzinski, J. S. Osinski, and P. D. Dapkus, *Appl. Phys. Lett.* **58**, 717 (1991).

<sup>7</sup>E. A. Fitzgerald, G. P. Watson, R. E. Proano, D. G. Ast, P. D. Kirchner, G. D. Pettit, and J. M. Woodall, *J. Appl. Phys.* **65**, 2220 (1989), and references therein.

<sup>8</sup>L. C. Kimerling and J. R. Patel, *Appl. Phys. Lett.* **34**, 73 (1979); G. L. Miller, D. Lang, and L. C. Kimerling, *Annu. Rev. Mater. Sci.* **7**, 377 (1977); L. C. Kimerling, *Mater. Res. Soc. Bull.* **16**, 42 (1991).

<sup>9</sup>A. Larsson and J. Maserjian, *Appl. Phys. Lett.* **58**, 1946 (1991).

<sup>10</sup>J. Maserjian, P. O. Andersson, B. R. Hancock, J. M. Iannelli, S. T. Eng, F. J. Grunthaner, K.-K. Law, P. O. Holtz, R. J. Simes, L. A. Coldren, A. C. Gossard, and J. L. Merz, *Appl. Opt.* **28**, 4801 (1989).

<sup>11</sup>D. H. Rich, A. Ksendzov, R. W. Terhune, F. J. Grunthaner, B. A. Wilson, H. Shen, M. Dutta, S. M. Vernon, and T. M. Dixon, *Phys. Rev. B* **43**, 6836 (1991).

<sup>12</sup>For migration in AlGaAs, see, e.g., S. Guha, A. Madhukar, K. Kaviani, L. Chen, R. Kuchibhotla, R. Kapre, M. Hyugaji, and Z. Xie, *Mater. Res. Soc. Proc.* **145**, 27 (1989).

<sup>13</sup>S. Luryi and E. Suhir, *Appl. Phys. Lett.* **49**, 140 (1986).

<sup>14</sup>M. J. Joyce, M. Gal, and J. Tann, *J. Appl. Phys.* **65**, 1377 (1989).

<sup>15</sup>P. M. Petroff, R. A. Logan, and A. Sauvage, *Phys. Rev. Lett.* **44**, 287 (1980); *J. Microsc.* **118**, 255 (1980).

<sup>16</sup>B. A. Wilson, C. E. Bonner, T. D. Harris, M. G. Lamont, R. C. Miller, S. K. Sputz, S. M. Vernon, V. E. Haven, R. M. Lum, and J. K. Klingert, *Mater. Res. Soc. Symp. Proc.* **91**, 255 (1987). Additionally, in a stress-induced deformation of Si and Ge samples, Schröter, Scheibe, and Schoen [*J. Microsc.* **118**, 23 (1980)] have argued that variations in the mobility and density of free holes, as determined from Hall measurements, are associated with clouds of point defects which surround dislocations.

<sup>17</sup>W. Hagen and H. Strunk, *Appl. Phys.* **17**, 85 (1978).

<sup>18</sup>J. W. Matthews, A. E. Blakeslee, and S. Mader, *Thin Solid Films* **33**, 253 (1976).

# Near-infrared cathodoluminescence imaging of defect distributions in $\text{In}_{0.2}\text{Ga}_{0.8}\text{As}/\text{GaAs}$ multiple quantum wells grown on prepatterned GaAs

D. H. Rich, K. C. Rajkumar, Li Chen,<sup>a)</sup> and A. Madhukar<sup>b)</sup>

Photonic Materials and Devices Laboratory, Department of Materials Science and Engineering,  
University of Southern California, Los Angeles, California 90089-0241

F. J. Grunthaner

Center for Space Microelectronics Technology, Jet Propulsion Laboratory,  
California Institute of Technology, Pasadena, California 91109

(Received 18 October 1991; accepted for publication 25 April 1992)

The defect distribution in a highly strained  $\text{In}_{0.2}\text{Ga}_{0.8}\text{As}/\text{GaAs}$  multiple-quantum-well (MQW) structure grown on a patterned GaAs substrate is examined with cathodoluminescence imaging and spectroscopy in the near infrared. By spatially correlating the luminescence arising from the MQW exciton recombination ( $\lambda \approx 950$  nm) with the longer wavelength ( $1000 \leq \lambda \leq 1200$  nm) luminescence arising from the defect-induced recombination, we demonstrate that it is possible to determine the regions of highest film quality in both the mesa and valley regions. The present approach enables a judicious determination of the optimal regions to be used for active pixels in InGaAs/GaAs spatial light modulators.

Presently, there is a considerable level of interest in using selective-area epitaxial growth approaches to achieve the fabrication of thick pseudomorphic films of strained  $\text{In}_x\text{Ga}_{1-x}\text{As}$  and related multiple-quantum-well (MQW) structures on GaAs substrates. Previous studies employing transmission electron microscopy (TEM),<sup>1-4</sup> photoluminescence (PL),<sup>2,5</sup> cathodoluminescence (CL),<sup>6</sup> optical absorption,<sup>1,3</sup> and micro-Raman<sup>4</sup> have demonstrated a substantial reduction in the density of misfit dislocations occurring in thick  $\text{In}_x\text{Ga}_{1-x}\text{As}$  films and MQWs grown on GaAs mesas with dimensions between 1 and 400  $\mu\text{m}$ . In order to optimize the procedure by which to fabricate functioning device arrays having such lateral dimensions, it is essential to be able to elucidate the optical and structural properties on the scale of a micrometer or less.

In this letter, we employ the technique of cathodoluminescence scanning electron microscopy (SEM) to examine spatial and spectral variations in the defect-induced near-infrared luminescence of a thick MQW structure of  $\text{In}_{0.2}\text{Ga}_{0.8}\text{As}/\text{GaAs}$  grown by molecular-beam epitaxy (MBE) on a prepatterned GaAs substrate. The present study demonstrates the complex interplay between the mesa side walls and the growth kinetics in the valley regions. These results are expected to influence current thinking regarding the design and fabrication of spatial light modulator device arrays requiring the highest possible material quality in isolated pixel regions.<sup>3,7</sup>

The  $n^+$ -type GaAs(100) substrate was patterned using conventional optical lithography and wet chemical etching.<sup>1-3</sup> The crystallographically selective etch results in undercut and inclined sidewall planes at the mesa edges. The resulting mesas had dimensions of about  $2 \mu\text{m} \times 16 \mu\text{m} \times 18 \mu\text{m}$  with a 40- $\mu\text{m}$  pitch in both directions before MBE growth. A SEM image of a typical mesa is shown in Fig. 1. The structure consists of a 100-period  $\text{In}_{0.2}\text{Ga}_{0.8}\text{As}$  (80 Å)/GaAs (160 Å) MQW (having ~2.38  $\mu\text{m}$  total

thickness) with a 5000-Å  $p^+$ -type GaAs capping layer. Cross-sectional TEM and absorption results of this sample (designated RG891110) have been previously described.<sup>1,3</sup> Cathodoluminescence measurements were performed at the Jet Propulsion Laboratory.<sup>8</sup> An electron beam current of 0.5 nA at an accelerating voltage of 40 kV was used to probe the sample for the CL measurements. The sample was maintained at a temperature of ~77 K.

A SEM image of the structure showing a  $220 \mu\text{m} \times 170 \mu\text{m}$  area is presented in Fig. 2(a). Scanning monochromatic CL images corresponding to the same region are shown in Figs. 2(b)–2(d) for wavelengths of 950, 1040, and 1120 nm, respectively. Regions of increasing luminescence signal are represented by areas of lighter shades. The image with  $\lambda = 950$  nm [in Fig. 2(b)] shows a hatched region of bright luminescence with maximum intensity in the valley regions midway along the [010] diagonal (dotted line in Fig. 1) between mesa centers. There is some observable emission from the centers and edges of the mesas at this wavelength. A dark halo region is found to surround the mesas, where little emission is detected. As the wavelength is increased to 1040 and 1120 nm, the images in Fig. 2 show an enhanced emission from the centers and sides of the mesas relative to the valley regions. Also, for these wavelengths the luminescence patterns on top of most of the mesas show a distinct dumbbell shape with maximal intensity near the mesa edges parallel to [110], as seen in Figs. 2(c) and 2(d).

In order to further quantify the spatial distribution of the luminescence, we have performed a set of line scans over the single mesa which is depicted in the SEM image of Fig. 1. In Fig. 3, the CL intensity for the three different wavelengths of 840, 950, and 1120 nm are plotted as a function of the electron-beam position ( $\Delta X$ ) along the dashed and solid lines in Fig. 1 which are oriented along [110] and [110], respectively. The arrows in Fig. 1 indicate an increasing  $\Delta X$  in the line scan which has its origin at the intersection of the dashed and solid lines; the point labeled

<sup>a)</sup>Also with Department of Physics, University of Southern California.

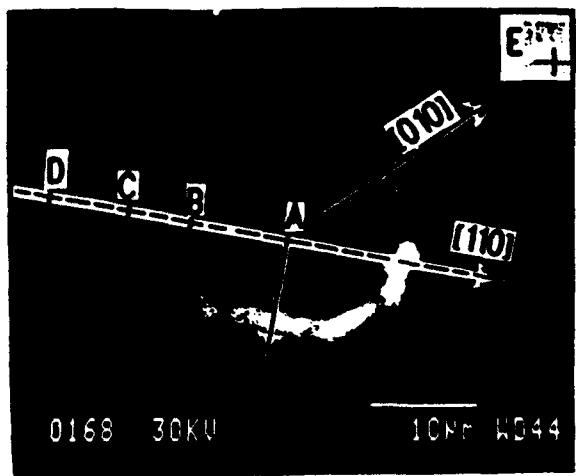


FIG. 1. SEM micrograph of a typical mesa of the  $\text{In}_{0.17}\text{Ga}_{0.83}\text{As}/\text{GaAs}$  MQW structure. The dashed and solid lines with arrows oriented along  $[110]$  and  $[010]$ , respectively, indicate the position of the electron beam during line scans. Points A-E indicate beam position during local CL spectroscopy.

A ( $\Delta X=0$ ) is located in the approximate center of the mesa.

These results were further investigated by localized CL spectroscopy of the region depicted in the SEM image of

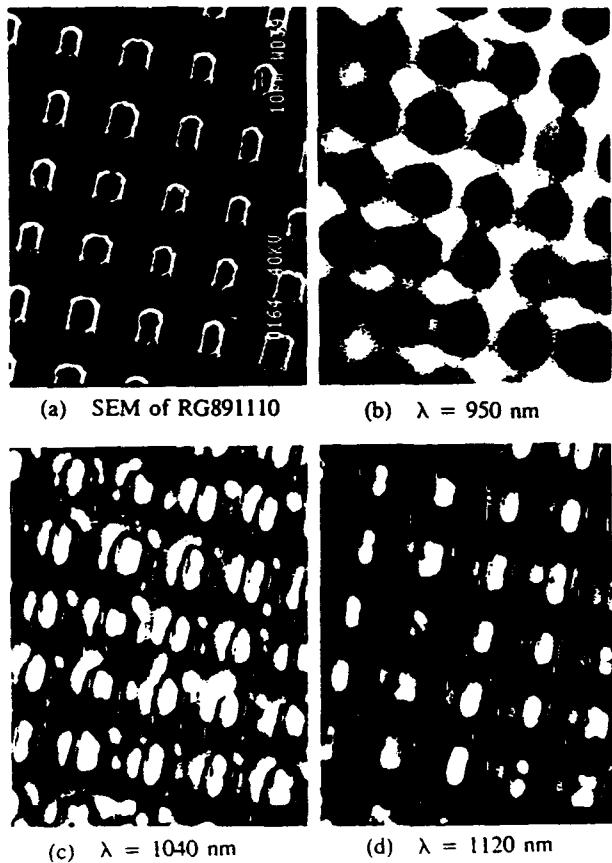


FIG. 2. SEM (a) and scanning monochromatic cathodoluminescence images [(b), (c), and (d)] correspond to wavelengths of 950, 1040, and 1120 nm, respectively] of the same region of the  $\text{In}_{0.17}\text{Ga}_{0.83}\text{As}/\text{GaAs}$  MQW structure.

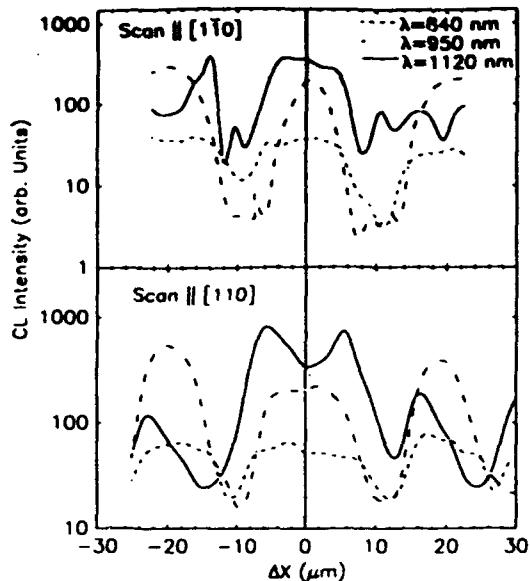


FIG. 3. Cathodoluminescence monochromatic line scans along the solid and dashed lines depicted in Fig. 1;  $\Delta X=0$  corresponds to point A in Fig. 1.

Fig. 1. The electron beam was fixed at certain points indicated in Fig. 1 with tick marks (labeled A, B, C, D, and E), and CL spectra, as presented in Fig. 4, were taken. Multiplication factors used to scale the absolute intensity of the spectra are indicated. The peak at  $\lambda=840$  nm is the result of emission from the GaAs capping layer. The sharp Gaussian-shaped peak at  $\sim 950$  nm is the MQW  $e-hh$  exciton transition which has a large variation in intensity dependent on the beam position. From the two line scans of Fig. 3, the  $\lambda=950$  nm peak is seen to maximize at positions of  $\Delta X \approx \pm 20 \mu\text{m}$  from the center of the mesa. Further inspection of the CL image of Fig. 2(b) and the local spectra in Fig. 4 at point E ( $\sim 28 \mu\text{m}$  from point A), however, indicates that the regions halfway along the  $[010]$  diagonal between mesa centers yield the strongest exciton peak intensity. A strong emission in the  $1000 \leq \lambda \leq 1200$  nm region (in Fig. 4) is observed for certain points A and B on the mesa; two broad bands centered at  $\lambda \approx 1040$  nm and  $\lambda \approx 1100$  nm are seen. Previously, from photoluminescence of single  $\text{In}_{0.17}\text{Ga}_{0.83}\text{As}$  QWs having various thicknesses, a long wavelength luminescence was found in the  $1000 \leq \lambda \leq 1600$  nm region.<sup>9</sup> This was attributed to presence of interface defects, such as misfit dislocations at the  $\text{InGaAs}/\text{GaAs}$  interface. Thus, it is reasonable to attribute the longer wavelength CL emission to the presence of structural defects which have previously been observed in cross-sectional TEM of this sample.<sup>1,3</sup> The difference in these two types of structures (single QW vs MQW and In alloy composition) would naturally give rise to differences in defect types and densities, and would therefore yield different luminescence spectra. The defect-induced luminescence is seen to be strongest near  $\Delta X \approx \pm 6 \mu\text{m}$  in the  $\lambda=1120$  nm line scan of Fig. 3, which is near the centers of dumbbell-shaped lobes seen in the CL images. The exciton emission is markedly reduced at point B ( $\Delta X=-9 \mu\text{m}$ )

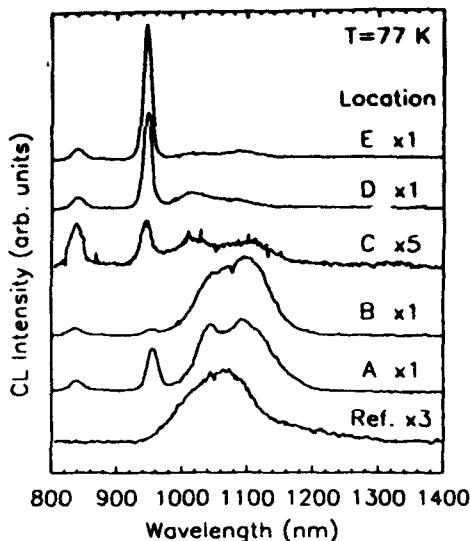


FIG. 4. Localized cathodoluminescence spectroscopy at the points A–E shown in Fig. 1. The sample labeled *Ref.* corresponds to the spectrum taken in the nonpatterned region of the  $\text{In}_{0.5}\text{Ga}_{0.5}\text{As}/\text{GaAs}$  MQW structure.

which is near the edge of the mesa in Fig. 1. This is consistent with enhancement of the defect-induced emission at point B since a large presence of defect-related recombination channels are expected to compete with the exciton recombination and reduce the total carrier lifetime.

The electron beam was further positioned within the dark halo region surrounding the mesa [see Fig. 2(b)] at point C ( $\Delta X = -15 \mu\text{m}$ ) in Fig. 1. A significant decrease in both the exciton- and defect-related features is seen. Thus, the presence of additional *nonradiative* recombination channels caused by changes in the defect distribution dominate the carrier lifetime in this region. The electron-beam positions D ( $\Delta X = -22 \mu\text{m}$ ) and E yield CL spectra with a large increase in the exciton luminescence and simultaneous reduction in the defect feature. Also at point E, the largest exciton-to-defect emission ratio was found; this indicates the region with the best-film quality. The spectrum labeled *Ref.* in Fig. 4 corresponds to a point on the nonpatterned region of the sample. Only a broad and shifted defect-related band is observed with no exciton emission. This is consistent with the marked reduction in the threading dislocation density in the patterned region relative to the nonpatterned region, as previously observed in cross-sectional TEM.<sup>1,3</sup>

The dumbbell shape on the mesas, as seen in the  $\lambda > 1000 \text{ nm}$  CL images and the [110] CL line scan, indicates the presence of interfacet cation migration in which the In migrates up along the inclined edges. The presence of contiguous low index planes is known to cause interfacet migration; this behavior has been observed to cause compositional variations in the AlGaAs and InGaAs films grown on patterned substrates.<sup>1–3,5,10</sup> The wavelength positions of the exciton peak at points A and E are 954 and 946 nm, respectively. If attributed exclusively to a variation in the In content, this shift corresponds to an In compositional change upper limit of  $\sim 0.5\%$ .<sup>11</sup> This is too

small a shift to alone account for the extreme carrier-induced luminescence variations observed here. The specific structural nature of the defects responsible for the long-wavelength luminescence cannot be determined from the present study. The defect-induced emission is, however, likely influenced by the proximity of the region to the growth on the contiguous (111) and higher index facets and the attendant interfacet cation migration. It is often desirable to eliminate the effects of cation migration on the faceted walls altogether by growing on elongated [110] striped mesas.<sup>2,5</sup>

In conclusion, we have demonstrated the effectiveness of cathodoluminescence imaging and spectroscopy in the near infrared in determining the spatial distribution of the defect-induced bands and exciton line for an InGaAs MQW grown on patterned GaAs. The presence of inclined mesa edges and cation migration is seen to influence the defect distribution for this particular class of patterned MBE growth. The presence of a characteristic dumbbell shape in the defect-induced luminescence pattern suggests that these regions are far from ideal for applications in spatial light modulators. Instead, regions midway between the mesa centers yield extremely large exciton-to-defect luminescence ratios. This suggests that these regions, with further device processing, could serve as *high-contrast-ratio* light modulators.<sup>3,7</sup>

At the University of Southern California, part of this work was supported by the Air Force Office of Scientific Research, University Research Initiative, Defense Advanced Research Projects Agency (DARPA), and Office of Naval Research. Other parts were performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, Caltech, and were jointly sponsored by DARPA and National Aeronautics and Space Administration, Office of Aeronautics, Exploration, and Technology.

- <sup>1</sup>A. Madhukar, K. C. Rajkumar, L. Chen, S. Guha, K. Kaviani, and R. Kapre, *Appl. Phys. Lett.* **57**, 2007 (1990).
- <sup>2</sup>S. Guha, A. Madhukar, and L. Chen, *Appl. Phys. Lett.* **56**, 2304 (1990).
- <sup>3</sup>L. Chen, K. Hu, K. C. Rajkumar, S. Guha, R. Kapre, and A. Madhukar, *Proceedings of the MRS Symposium on Materials for Optical Information Processing*, Anaheim, CA, Vol. 240, May 1–3, 1991 (in press).
- <sup>4</sup>W. C. Tang, H. J. Rosen, S. Guha, and A. Madhukar, *Appl. Phys. Lett.* **58**, 1644 (1991).
- <sup>5</sup>Y. Zou, P. Grodzinski, J. S. Osinski, and P. D. Dapkus, *Appl. Phys. Lett.* **58**, 717 (1991).
- <sup>6</sup>E. A. Fitzgerald, G. P. Watson, R. E. Proano, and D. G. Ast, P. D. Kirchner, G. D. Pettit, and J. M. Woodall, *J. Appl. Phys.* **65**, 2220 (1989).
- <sup>7</sup>K. Hu, L. Chen, A. Madhukar, P. Chen, K. C. Rajkumar, K. Kaviani, Z. Karim, C. Kyriakakis, and A. R. Tanguay, Jr., *Appl. Phys. Lett.* **59**, 1108 (1991), and references therein.
- <sup>8</sup>D. H. Rich, A. Ksendzov, R. W. Terhune, F. J. Grunthaner, B. A. Wilson, H. Shen, M. Dutta, S. M. Vernon, and T. M. Dixon, *Phys. Rev. B* **43**, 6836 (1991).
- <sup>9</sup>M. J. Joyce, M. Gal, and J. Tann, *J. Appl. Phys.* **65**, 1377 (1989).
- <sup>10</sup>For migration in AlGaAs, see, e.g., S. Guha, A. Madhukar, K. Kaviani, L. Chen, R. Kuchibhotla, R. Kapre, M. Hyugaji, and Z. Xie, *Mater. Res. Soc. Proc.* **145**, 27 (1989).
- <sup>11</sup>K. F. Huang, K. Tai, S. N. G. Chu, and A. Y. Cho, *Appl. Phys. Lett.* **54**, 2026 (1989).

# Analysis of weighted fan-out/fan-in volume holographic optical interconnections

Praveen Asthana, Gregory P. Nordin, Armand R. Tanguay, Jr., and B. Keith Jenkins

The feasibility of employing volume holographic techniques for the implementation of highly multiplexed weighted fan-out/fan-in interconnections is analyzed on the basis of interconnection fidelity, optical throughput, and complexity of recording schedule or implementation hardware. These feasibility criteria were quantitatively evaluated using the optical beam propagation method to numerically simulate the diffraction characteristics of volume holographic interconnections recorded in a linear holographic material. We find that conventional interconnection architectures (that are based on a single coherent optical source) exhibit a direct trade-off between interconnection fidelity and optical throughput on the one hand, and recording schedule or hardware complexity on the other. In order to circumvent this trade-off we describe and analyze in detail an incoherent/coherent double angularly multiplexed interconnection architecture that is based on the use of multiple-source array of individually coherent but mutually incoherent sources. This architecture either minimizes or avoids several key sources of cross talk, permits simultaneous recording of interconnection weights or weight updates, and provides enhanced fidelity, interchannel isolation, and throughput performance.

## 1. Introduction

Volume holographic optical elements (VHOE's) have often been suggested as the principal components of an optical interconnection technology for applications such as optical computing and telecommunications that require a large number of interconnections.<sup>1-6</sup> Depending on the application, such interconnection systems may also require varying degrees of interconnection weighting, fan-out, fan-in, and channel independence. In particular, interconnection systems for artificial neural networks provide a specific application for which all of these issues are important.

Artificial neural networks are composed of many highly interconnected nonlinear computational elements (neuron units) that operate in parallel and are arranged in architectural patterns that are motivated to a certain extent by investigations of biological

neural networks. The computational elements are in most cases densely interconnected with weighted connection pathways that can be reconfigured and updated to permit either supervised or unsupervised learning. Implementations of adaptive neural networks should optimally permit such pathway reconfiguration and weight updates without excessively compromising either hardware complexity or computational efficiency.

Typical artificial neural-network interconnection topologies require a high degree of both fan-out and fan-in at each neuron unit. A fully interconnected topology for the case of a single-layer network is illustrated schematically in Fig. 1, in which two planes of neuron units are shown. In such a fully connected network, the output of each neuron unit in the input plane is fanned out to all of the neuron units in the output plane. Similarly, each neuron unit in the output plane receives the weighted fan-in of all of the neuron-unit outputs from the input plane. The fan-out/fan-in requirements of neural networks act as a multiplier on the total number of interconnected neuron units, resulting in an interconnection system that must prove capable of supporting very large numbers of independent pathways in a relatively compact topology. For example, a fully connected neural network having  $10^5$  neuron units in both the input and output planes requires  $10^{10}$  interconnections, as does a partially connected neural network

---

When this work was performed, the authors were with the Optical Materials and Devices Laboratory, Center for Photonic Technology, Signal and Image Processing Institute, and the Center for Neural Engineering, University of Southern California, University Park, MC-0483, Los Angeles, California 90089-0483. P. Asthana is currently with the Storage Systems Products Division, IBM Corporation, Tucson, Arizona 85744; G. P. Nordin is currently with the Electrical and Computer Engineering Department, University of Alabama in Huntsville, Huntsville, Alabama 35899.

Received 16 July 1992.

0003-6935/93/081441-29\$05.00/0

© 1993 Optical Society of America.

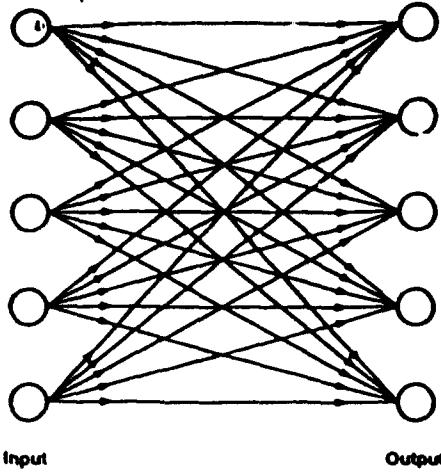


Fig. 1. Schematic representation of fan-out/fan-in interconnections between input and output planes of neuron units.

having  $10^6$  neuron units in both the input and output planes with a fan-out and fan-in of  $10^4$ .

The use of volume holographic optical elements has been proposed as the basis of an interconnection technology for neural networks precisely because it has the potential to meet the critical requirements of providing both large numbers of interconnections and a weighted fan-out/fan-in topology.<sup>7</sup> Even so, the feasibility of using VHOE's for large-scale weighted fan-out/fan-in interconnection applications depends further on the fidelity with which the interconnection weights can be implemented and on the optical throughput that can be achieved in the volume holographic interconnection system. The optical throughput is a measure of the fraction of the total incident optical power that is diffracted into the set of desired outputs; as such, it provides a quantitative assessment of the overall interconnection system efficiency. Important additional implementation issues include the total number of required exposures, the optical power incident on the holographic recording medium per exposure, the complexity of the implementation hardware, the required exposure schedule (if any), and the total recording time for a given interconnection complexity.

In this paper we quantitatively evaluate the performance characteristics of two distinct volume holographic interconnection architectures in accordance with these criteria. The first interconnection architecture is novel in that an array of individually coherent but mutually incoherent optical sources is used to generate a multiplicity of angularly multiplexed recording beams.<sup>8-14</sup> We hereinafter refer to this type of interconnection architecture as being incoherent/coherent double angularly multiplexed. The second interconnection architecture is based on the use of a single coherent optical source during recording, as has been widely discussed in the literature.<sup>3,15-19</sup> This conventional architecture is referred to herein as a single-source architecture.

In our analysis we use the optical beam propagation method (BPM)<sup>14,20-23</sup> to investigate the diffrac-

tion characteristics of weighted fan-out/fan-in interconnections in complex systems that involve both large numbers of holographic gratings and multiple readout beams, and that therefore do not lend themselves readily to analytical solutions. For our purposes herein, we consider explicitly the case of linear holographic materials in order to illustrate key differences in interconnection performance that are architecture dependent. Generalization of these results to nonlinear holographic materials (for example, to certain photorefractive media) is beyond the scope of this paper.

Results are reported herein on the numerical simulation of up to 10- (input node) to-10 (output node) weighted fan-out/fan-in volume holographic interconnections that incorporate between 10 and 190 individual holographic gratings multiplexed within the same region of the volume holographic recording medium, depending on the specific architectural configuration considered. Such simulations require significant run times, even when implemented with the highly efficient BPM algorithm on a supercomputer. To the best of our knowledge, the 10-to-10 case with fully independent weights is one of the most complex volume holographic interconnection systems that have been analyzed to date.

Our simulations demonstrate that the novel incoherent/coherent double angularly multiplexed architecture exhibits both high interconnection fidelity and high optical throughput efficiency even in the presence of fully simultaneous recording. This combination of desirable characteristics derives from the elimination (or minimization) of several distinct sources of interchannel cross talk and throughput loss that are unavoidably present in the more widely investigated single-source architecture. In addition, the mutual incoherence of the readout beams in the incoherent/coherent double angularly multiplexed architecture provides naturally for linearity of summation in an intensity representation without an associated fan-in loss.

In order to compare the fidelity and throughput performance of the incoherent/coherent double angularly multiplexed architecture with an appropriate benchmark, we analyzed the single-source architecture (illustrated in Fig. 2) under directly comparable holographic recording conditions. In particular, we examined the cases of simultaneous, pagewise-sequential, and fully sequential recording of the desired weight updates. These three recording methods represent distinct trade-offs between interchannel cross talk resulting from the presence of extraneous gratings, on the one hand, and recording schedule or hardware complexity on the other. The presence of such extraneous gratings can cause significant errors in the diffracted outputs that are fanned in to a given interconnection node. For example, in the 10-to-10 single-source interconnection system that we modeled, errors as large as 100% in the relative diffracted outputs occur at the peak optical throughput of 50% for the simultaneous recording method. To the best of our knowledge, the analysis of the

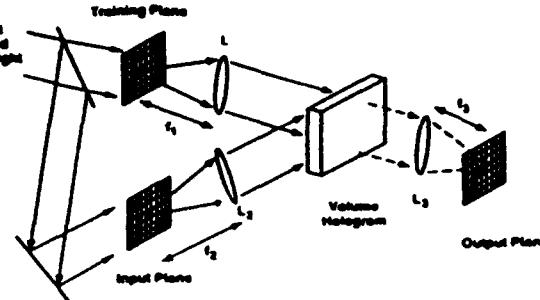


Fig. 2. Schematic diagram of a single-source holographic interconnection architecture in which diffraction gratings in a volume hologram connect pixels in the input plane to pixels in the output plane. Interconnection gratings are formed by recording the interference between light from pixels in the training plane and light from pixels in the input plane.  $L_1-L_3$  are lenses;  $f_1-f_3$  are focal lengths.

simultaneous recording method (as applied to the single-source architecture) presented herein establishes the fidelity errors and throughput losses of this technique quantitatively for the first time. Elimination of the extraneous gratings by using the fully sequential recording method (at a substantial cost in recording schedule or hardware complexity) reduces the largest errors to 13% at an increased peak optical throughput of 85%. These errors for the fully sequential recording case are significantly lower than those described in Ref. 19, at least in part because the relative phase relationships among the recording beams are maintained on readout in our simulations.

In addition, we demonstrate that the severe fan-in loss that unavoidably accompanies the collinear combination of mutually incoherent beams of essentially common wavelength,<sup>24</sup> and that also applies to incoherent readout in the single-source architecture, is one of several effects of beam degeneracy (defined in Section 4.D).<sup>9,12,13,25</sup> This same physical mechanism is also primarily responsible for the residual errors apparent in the fully sequential recording method at peak throughput.

The remainder of this paper is organized as follows. Specific aspects of volume holographic interconnection systems that are essential for the establishment of a valid comparison metric are discussed in Section 2, including the choice of signal representations and weights appropriate for photonic neural networks, the implementation of weighted fan-out/fan-in interconnections in a single-source architecture, and the associated sources of fidelity errors and throughput losses. Section 3 details our modeling methodology and assumptions, anticipating the discussion of the simulation results obtained for the single-source architecture that are provided in Section 4. We present two distinct configurations of the incoherent/coherent double angularly multiplexed architecture and quantitatively analyze their performance characteristics in Section 5. Finally, we compare the relative advantages and disadvantages of each architectural configuration and summarize our results in Section 6.

## 2. Preliminary Concepts

### A. Basic Neural-Network Operation

In many neural networks the computational process of a single network layer can be represented by

$$y_i = f(p_i), \quad (1)$$

$$p_i = \sum_{j=1}^N W_{ij} x_j, \quad (2)$$

in which neuron units  $i$  and  $j$  are in the output and input planes, respectively,  $y_i$  is (proportional to) the output of neuron unit  $i$ ,  $x_j$  is (proportional to) the output of neuron unit  $j$ ,  $W_{ij}$  is the weight of the interconnection between neuron units  $i$  and  $j$ ,  $N$  is the number of neurons in the input plane,  $f$  is the nonlinear threshold function of each neuron unit, and  $p_i$  is the activation potential.

In a neural network that incorporates learning, the weights  $W_{ij}$  are updated incrementally according to an appropriate learning rule as training patterns are presented to the network. A large class of learning algorithms can be characterized by the following weight-update rule:

$$\Delta W_{ij} = \alpha \delta_i^{(m)} x_j^{(m)}, \quad (3)$$

in which  $\Delta W_{ij} = W_{ij}(m) - W_{ij}(m-1)$  is the weight update,  $\alpha$  is the learning gain constant, and  $m$  is the iteration index. Various learning rules can be formed by suitably choosing the functional form of  $\delta_i^{(m)}$ ; for example, in the case of Hebbian learning,  $\delta_i^{(m)}$  is chosen to be  $y_i$ .<sup>26</sup>

### B. Weighted Fan-Out/Fan-In Interconnections in a Single-Source Architecture

The conventional single-source volume holographic interconnection architecture as configured for photonic neural-network implementation is shown schematically in Fig. 2. This architecture consists of a holographic interconnection medium and three planes of neuron units: a training plane, an input plane, and an output plane. The two-dimensional (2-D) arrays of neuron units or pixels (since in many photonic neural-network implementations each pixel corresponds to a single neuron unit) are placed on these planes. In applications other than neural networks these pixels can represent generic connection nodes. The arrays of pixels can be implemented using, for example, 2-D spatial light modulators (SLM's). Both the training- and input-plane SLM's are illuminated by a single coherent source in the configuration shown in Fig. 2.

A given weighted interconnection between a pixel in the input plane and a pixel in the output plane is physically realized as a single diffraction grating in the holographic medium. Each grating is formed by recording the interference pattern generated by coherent superposition of light from a pixel in the input plane and light from a pixel in the training plane. For simplicity the pixels in both planes are typically

sumed to act as point source<sup>3,18</sup> and lenses L<sub>i</sub> and L<sub>j</sub> function as collimating lenses. Light from each pixel is thus incident as an angularly distinct collimated beam (typically approximated as a plane wave<sup>3,18</sup>) on the holographic medium. The light from each pixel is assumed to fully illuminate the aperture of the holographic medium to ensure the potential for full connectivity.<sup>3,18</sup>

**2. Signal Representation in a Single-Source Architecture**  
Depending on the computational algorithm, the optical signal from each input pixel (after collimation) can be represented by a number of different physical quantities, including intensity, amplitude, polarization, and wavelength. In the two most commonly investigated approaches<sup>3</sup> the signal is represented either by the complex electric field or by the intensity of the light transmitted or reflected by the pixel depending on the type of SLM).

In the amplitude representation the electric field of the light from pixel *j* that is incident at the front face of the hologram can be expressed (neglecting the vector nature of the electric field) as  $x_j^{\text{amp}} E_0 \exp(i\mathbf{k} \cdot \mathbf{r} + i\phi_j - i\omega t)$ , in which  $x_j^{\text{amp}}$  is the amplitude transmissivity (or reflectivity) of the *j*th pixel,  $E_0$  is the magnitude of the electric field amplitude of the readout beam,  $\phi_j$  is the phase of the readout beam (relative to an arbitrarily chosen coordinate system), and  $\omega$  is the angular frequency. In this representation a given interconnection weight is proportional to the amplitude diffraction efficiency of the corresponding interconnection grating.

In the intensity representation the output of pixel *j* is expressed as  $x_j I_0$ , in which  $x_j$  is the intensity transmittance (or reflectance) of the SLM pixel and  $I_0 = E_0^2$ . In this case each weight is proportional to the intensity diffraction efficiency of its corresponding interconnection grating.

When readout is performed with beams from more than one input pixel, the intensity detected within each output pixel consists of a weighted sum of diffracted signals. If an optical system is constructed such that the light beams from the input pixels are mutually incoherent during readout, the relative detected intensity (using the intensity representation) within the *i*th output pixel  $p_i$  is

$$p_i = \sum_{j=1}^N W_{ij} x_j, \quad (4)$$

in which  $W_{ij}$  is an intensity weight (which is proportional to the diffraction efficiency of grating *ij*).<sup>17</sup> The relationship expressed in Eq. (4) is the same as that in Eq. (2), which defines how the activation potentials are related to the inputs by the interconnection weights of the neural network. Thus readout with mutually incoherent beams in conjunction with an intensity representation leads to an optical system that performs the desired neural interconnection function of Eq. (2).

If the optical system is constructed such that the

light beams from the input-plane pixels are mutually coherent (as is typically the case for a single-source architecture), the diffracted signals detected in the output plane can be written (assuming an intensity representation) as

$$p_i = \left| \sum_{j=1}^N (W_{ij} x_j)^{1/2} \exp[i(\phi_j + \Phi_{ij})] \right|^2, \quad (5)$$

in which  $\Phi_{ij}$  is the phase of the *ij*th grating. Each grating phase is set by the relative phases of the recording beams used to form the grating and the properties of the holographic recording medium. If the argument of the exponential is constant for all *i* and *j*, the diffracted outputs for a single-source architecture using the intensity representation reduce to

$$p_i = \left| \sum_{j=1}^N (W_{ij} x_j)^{1/2} \right|^2. \quad (6)$$

This equation is valid only if two specific conditions are met: (1) the readout beams must have the same relative phases as were used during recording of the interconnection gratings, and (2) the phase shift induced by the holographic recording medium itself must be constant for all recorded gratings.

Even with these assumptions, the square-root relationship embodied in Eq. (6) deviates substantially from the desired interconnection function of Eq. (2). The general effect of this deviation on neural-network operation has not yet been established.

As a result, several authors have chosen an amplitude representation for use in the single-source architecture, such that the diffracted outputs are given by<sup>3,27</sup>

$$p_i = \left| \sum_{j=1}^N W_{ij}^{amp} x_j^{amp} \exp[i(\phi_j + \Phi_{ij})] \right|^2, \quad (7)$$

in which  $W_{ij}^{amp}$  is the amplitude weight (which is proportional to the amplitude diffraction efficiency). If the argument of the exponential in Eq. (7) is constant for all *i* and *j*, the system yields the square of the desired interconnection function, which may be corrected either electronically or by adjusting the nonlinear threshold function of each neuron unit. This simplification is dependent, as is Eq. (6), on the two conditions specified above. The first of these conditions is contingent on the maintenance of rigid optical phase stability in the system, which may be difficult to realize in practice.

A primary advantage of readout with mutually incoherent beams is the avoidance of the rigid phase-stability requirements that are necessary for coherently read out systems.<sup>3,24,26</sup> However, a serious disadvantage associated with the use of mutually incoherent readout beams is the significant throughput loss that is characteristic of all holographic interconnection architectures in which the fan-in is performed collinearly.<sup>24</sup> Since single-source intercon-

nection systems rely on collinear fan-in, they inherently suffer from this throughput loss for readout with mutually incoherent beams. In Section 5 we discuss a method for using readout with mutually incoherent beams without suffering an incoherent fan-in loss.

A further consideration in the implementation of photonic neural networks is that many neural algorithms require the use of bipolar weights and bipolar neuron-unit outputs. One method of achieving this bipolarity for readout with mutually coherent beams is to use an amplitude representation in which a 180° phase shift in the phase of both the electric fields and the diffraction gratings is used to represent negative numbers.<sup>3,27</sup> While attractive in principle (requiring only a single data channel per neuron unit), this approach is difficult to implement in practice, in part because the phases of the resultant diffracted outputs must be detected as well as their amplitudes. An alternative method of achieving bipolarity is to use an intensity representation in conjunction with a dual-rail concept in which each neuron has separate positive and negative channels for both input and output signals with associated weighted interconnections for each pair of channels.<sup>13,27,29,30</sup> Although this method requires two data channels per neuron unit, it is compatible with simple square-law detectors and does not require mutually coherent readout beams.

For our purposes herein, we consider only unipolar weights and unipolar neural outputs since the dual-rail method can be used to generalize to the fully bipolar case. In addition, we choose to adopt the intensity representation throughout in order to facilitate direct comparisons between mutually coherent and mutually incoherent readout systems. In particular, the intensity representation is perhaps the most natural representation for the incoherent/coherent double angularly multiplexed interconnection architecture, yielding linear sum rules in the diffracted outputs. In any case, this choice does not affect the general conclusions drawn regarding interconnection fidelity and optical throughput. For the various cases of mutually coherent readout treated herein, we assume that both conditions (1) and (2) specified above [following Eq. (6)] are met.

#### D. Recording Methods in a Single-Source Architecture

For the single-source architecture shown in Fig. 2, in which the light from each pixel is mutually coherent with the light from all such pixels in a given plane, there are several methods by which the desired interconnection gratings may be recorded. All such methods involve a sequence of exposures, and they differ in the nature of each individual exposure. The first method is simultaneous recording, in which mutually coherent light beams from all the input and training pixels are incident simultaneously on the holographic medium during each exposure. By recording the resultant interference pattern, the volume holographic medium forms gratings that con-

nect each neuron unit in the input plane with each neuron unit in the output plane. These constitute the desired *intra-planar* connections that perform the weighted fan-out/fan-in function. However, *intra-planar* gratings that connect pairs of pixels within the input plane (and within the training plane as well) are formed as a result of the mutual coherence of the sources.<sup>16,17</sup> These extraneous gratings (termed coherent-recording cross-talk gratings, or cross gratings) introduce a serious source of cross talk into the interconnection system that in turn leads to a loss of both throughput and reconstruction fidelity.

In the pagewise-sequential recording method, a single pixel in the input plane is connected simultaneously with all of its associated training pixels during each exposure, which eliminates the coherent cross-talk gratings within the input plane only. In the fully sequential recording method, the remaining coherent cross-talk gratings in the training plane are eliminated by recording connections between a single pixel in the input plane and a single pixel in the training plane during each exposure. In order to achieve the reduction in coherent-recording cross talk offered by these methods, however, an increase in the recording schedule or hardware complexity must be accommodated. In certain photorefractive media the situation may be complicated further by the need for a complex recording schedule to compensate for the partial erasure of previously recorded interconnection gratings during the recording of later gratings.<sup>3,31</sup>

In a photonic neural-network implementation, the interconnection weights can be either precomputed (in a photonic or electronic computing system) and stored in a permanent holographic medium for later use or obtained adaptively using a suitable learning algorithm and a dynamic holographic medium. For the recording of precomputed weights it is clearly advantageous to minimize the total number of exposures, the total recording time, and the total exposure energy. For a system in which  $N$  input nodes are connected to  $N$  output nodes (referred to herein as an  $N$ -to- $N$  interconnection system) the minimum number of exposures required to record a set of fully independent weights is equal to  $N$ , as can be derived from degrees-of-freedom considerations.<sup>32</sup> The assumption of fully independent weights is equivalent to the assumption that the weight matrix is of rank  $N$ . If the full set of precomputed weights is recorded pagewise sequentially, therefore, the minimum number of exposures can be achieved.

For adaptive computation of weights a set of training pairs is presented sequentially (one training pair at a time) to the network. Each training pair consists of an input image and its corresponding training image, which are presented on the respective input and training planes of Fig. 2. The  $m$ th input image can be represented by the vector  $\mathbf{x}^{(m)}$ , the components  $x_i^{(m)}$  of which are shown in the first term on the right-hand side of Eq. (3). Similarly, the  $m$ th training image is represented by the vector  $\delta^{(m)}$ . Ideally,

le full input and training images for the  $m$ th training pair  $[x^{(m)}, \delta^{(m)}]$  are presented simultaneously at the input and training planes, respectively, such that only one exposure is required for each training pair. The outer product of the vectors  $x^{(m)}$  and  $\delta^{(m)}$  is recorded in each such exposure, which thus corresponds to the simultaneous recording case discussed above. However, if a pagewise-sequential or fully sequential recording strategy is necessary to avoid the deleterious effects of coherent-recording cross talk, then the number of exposures per training pair becomes  $N$  and  $N^2$ , respectively, for an  $N$ -to- $N$  interconnection system. If  $M$  training pairs are required to fully train the network, the total number of exposures for the simultaneous, pagewise-sequential, and fully sequential recording methods are  $M$ ,  $NM$ , and  $N^2M$ , respectively. If real-time adaptation is not required and precomputation of weights is permitted, then the minimum numbers of exposures required for the three recording methods are  $N$ ,  $N$ , and  $N^2$ , respectively.

The choice of recording method has significant practical consequences for photonic neural-network implementations (using the single-source interconnection architecture) in which adaptive computation of the weights is desired. As an example, if the SLM frame time is the temporal bottleneck of the system which represents perhaps a worst-case estimate, in that the single-pixel access time may be considerably shorter than the full SLM frame time for certain SLM's), then the total amount of time required to train the network scales linearly with the number of required exposures. For large numbers of interconnections (exactly the situation for which holographic interconnections are presumably attractive) this could result in impractically long training sessions. For example, if  $N = 10^4$ ,  $M = 10^3$ , the SLM's support a 10-ms update rate (such as for nonferroelectric liquid-crystal-based SLM's), and the holographic-material response time at the available power level is fast enough to not provide an even stricter bound, then the simultaneous, pagewise-sequential, and fully sequential recording methods would require 10 s,  $10^5$  s (28 h), and  $10^9$  s (32 years), respectively, to record the desired interconnections.

In the pagewise-sequential recording method, one technique for avoiding prolonged training sessions and consequently inefficient use of available optical power) is to focus the full incident beam on a given pixel of the input SLM and to scan the beam pixel-by-pixel during the recording of a given training pair. If scanning all of the input-plane pixels can be accomplished within the frame time of the SLM which corresponds to a 1- $\mu$ s dwell time for the above example), then pagewise-sequential recording can be accomplished in the same total time as simultaneous recording. The concept can also be applied to the fully sequential recording method (which requires 100-ps dwell times for the example above). Decreases in the total training time (with concomitant increases in the efficiency of power and energy use) for both

recording methods are achieved at the cost of additional system complexity, particularly since in many cases provision must also be made for simultaneous readout of all input-plane pixels during posttraining computation.

From the joint perspectives of recording schedule and hardware complexity, simultaneous recording may prove to be the most desirable recording method (in the adaptive neural-network paradigm) because the full parallelism of the optical architecture is used, thereby achieving the greatest computational throughput during training without resorting to additional components that increase the system complexity. However, simultaneous recording within the single-source architecture is perhaps the least desirable recording method from the perspective of interconnection fidelity resulting from the deleterious effects of coherent-recording cross talk.

### E. Sources of Fidelity Errors and Throughput Losses

In addition to coherent-recording cross talk, there are at least two other sources of fidelity errors and throughput losses that can be present in multiplexed fan-out/collinear fan-in interconnections implemented within the single-source architecture. The first of these is grating-degeneracy cross talk.<sup>16-18,33,34</sup> This form of cross talk arises because of the particular geometric placement of the recording pixels used. If the pixels are placed on regular grids, gratings with degenerate wave vectors may be recorded. Desired interconnections that have degenerate grating wave vectors can have severely distorted weights during reconstruction. This form of cross talk may be alleviated by placing the pixels on fractal sampling grids.<sup>34</sup> The cost, however, is the need to subsample the input, training, and output planes, which decreases the interconnection density of the system.

The second form of cross talk is beam-degeneracy cross talk, which arises from degeneracies in the wave vectors of beams diffracted from different gratings.<sup>9,12,13,25</sup> This form of cross talk is inherent in fan-out/fan-in volume holographic interconnection systems in which the diffracted beams that constitute a given fan-in exit collinearly from the holographic medium; it is present regardless of the sampling grids used for the input and training planes. In such interconnection systems, beam-degeneracy cross talk is also present even in the absence of cross gratings. For readout with mutually coherent beams in the single-source architecture, our simulation results indicate that beam degeneracy is a significant source of fidelity error only when the gratings are overmodulated. Further simulation results indicate that beam degeneracy is the primary physical mechanism responsible for the incoherent fan-in loss observed in single-source architectures when readout is performed with mutually incoherent beams.<sup>12,13,25</sup>

Related effects that can be attributed to beam degeneracy have been observed for 2-to-1 beam combining using a coupled-wave analysis.<sup>35,36</sup> In addition, Lee *et al.* have estimated the magnitude of

degeneracy cross talk (identified therein as "order cross talk) insofar as it affects reconstruction fidelity in a limiting case.<sup>17</sup> Similarly, Slinger mentions that beam degeneracy (identified therein as one of several forms of multiple-grating interactions) is a potential source of cross talk.<sup>19</sup>

### 3. Modeling Methodology and Assumptions

In this section we describe our modeling methodology and assumptions in detail. The motivation behind our choice of modeling technique (the optical beam-propagation method, or BPM) is discussed in Section 3.A, followed by a brief outline of the pertinent features of the BPM algorithm in Section 3.B. In Section 3.C we discuss the geometric dimensions of the single-source interconnection architecture that we modeled, the recording characteristics of the holographic medium, and the method used to obtain the weight matrices for the various cases.

#### A. Choice of Simulation Method

The analysis of the diffraction properties of a large number of weighted gratings multiplexed in a volume holographic material has been a difficult problem. Several techniques have been used in the past to analyze multiple-grating diffraction.<sup>19,36-42</sup> Of these techniques, coupled-wave theory has been the most extensively used.<sup>19,36-40</sup> Application of this method often involves making a number of simplifying assumptions so that analytical or numerical solutions are more easily obtainable. Such assumptions include the use of a 2-D model (in which the gratings and the incident and diffracted beams all lie in a plane) and TE polarization of the beams.<sup>19,36-39</sup> With these assumptions and the additional assumption of no undesired cross gratings, analytical solutions have been obtained for the cases of 1-to- $N$  weighted fan-outs,<sup>38</sup>  $N$ -to-1 weighted fan-ins,<sup>39</sup> and  $N$ -to- $N$  weighted fan-out/fan-ins.<sup>19</sup> In Ref. 19 the assumption was further made that the weight matrix had a rank of one, which is achieved in practice in a linear holographic medium by recording each  $x_i$  with the same training vector,  $\delta$  (i.e., only a single training pair is recorded). For weight matrices having a higher rank (which is the relevant situation for photonic neural-network implementations) the system of coupled differential equations obtained using coupled-wave analysis has not yielded an analytical solution and thus must be solved numerically.<sup>19</sup> For example, Slinger has numerically modeled a 5-to-5 interconnection system for pagewise-sequential and fully sequential recording of the desired interconnection gratings.<sup>19</sup> However, numerical solutions for the simultaneous recording case have not been presented in the literature to our knowledge.

Rather than obtaining numerical solutions based on the results of a coupled-wave analysis, we used the optical beam propagation method<sup>20</sup> to numerically model readout of a volume hologram in which multiple interconnection gratings are stored. An advantage of this approach is that fewer simplifying as-

sumptions are necessary to make the problem computationally tractable. In coupled-wave analysis, for example, only Bragg-matched interactions between the readout (and diffracted) beams and the recorded gratings are typically retained. In Ref. 19 this led to the assumption that all undesired cross gratings for pagewise-sequential recording operate in the Bragg regime. However, in many physical geometries (depending on the incidence angles of the writing beams and the thickness of the holographic recording material) some or all of these cross gratings may actually operate in the Raman-Nath diffraction regime or have properties that are in the transition regime between Raman-Nath and Bragg diffraction (hereinafter referred to as the transition regime). For example, the interconnection geometry that we simulate (the dimensions of which are discussed in Section 3.C.1) has cross gratings that operate in all three of the possible diffraction regimes.

Coupled-wave theory could be used to model cross gratings that operate in the Raman-Nath or transition diffraction regimes if multiple diffraction orders for each cross grating are retained in the calculations. However, BPM has proved to be a significantly faster numerical technique for solving diffraction grating problems involving many spectral orders when the same number of orders are considered and the same level of accuracy is required.<sup>23</sup>

Restricting the analysis to only Bragg-matched interactions using the coupled-wave approach also neglects a further possible source of cross talk, namely, diffraction from non-Bragg-matched interconnection gratings. By way of contrast, diffraction from non-Bragg-matched gratings is incorporated directly in the BPM algorithm. Although the effects of such non-Bragg-matched interconnection gratings are small, we show in Section 4.E that diffraction from such gratings is the limiting source of fidelity error as the strength of the interconnections in a single-source architecture approaches zero.

#### B. Optical Beam Propagation Method

The motivation for our choice of computational method is discussed above; herein we give a brief qualitative overview of the BPM algorithm. References 20-23 include discussions of the derivation and/or the validity of BPM for various diffraction problems. The BPM analysis discussed herein considers two physical dimensions [the nominal propagation ( $z$  axis) and transverse ( $x$  axis) dimensions shown in Fig. 3] and assumes TE polarization for convenience.

The optical beam propagation method simplifies the analysis of volume grating diffraction by replacing the physically distributed modulation/diffraction problem with a sequence of infinitesimally thin modulation layers separated by homogeneous regions in which only diffraction occurs. If a sufficient number of modulation layers are incorporated in the calculation, the system can closely approximate the characteristics of a volume hologram. The calculation

caused by reductions in the grating modulation depth during recording.<sup>12</sup>

Regardless of the practicality of pagewise-sequential recording, our modeling of a 10-to-10 single-source architecture shows that if the beam splitter ratio is made large enough to reduce the magnitudes of the cross gratings relative to the interconnection gratings, then the fidelity and throughput of the interconnection system can be improved as compared with the simultaneous-recording method. For example, Fig. 7 shows the diffracted outputs, ratios, and ratio percentage errors for readout with mutually coherent beams. A beam splitter ratio of 1000 is assumed, which adjusts the average strength of the interconnection gratings to approximately three times the strength of the cross-talk gratings. As seen in Fig. 7(c), the percentage errors of the ratios show a marked improvement over the simultaneous-recording case. However, at the peak throughput of nearly 80% (see Fig. 6), the largest percentage errors in the ratios are still 50%. At 10% throughput the largest error is ~15%. As shown in Section 4.C, the fidelity can be further improved when the cross gratings are eliminated entirely by using fully sequential recording.

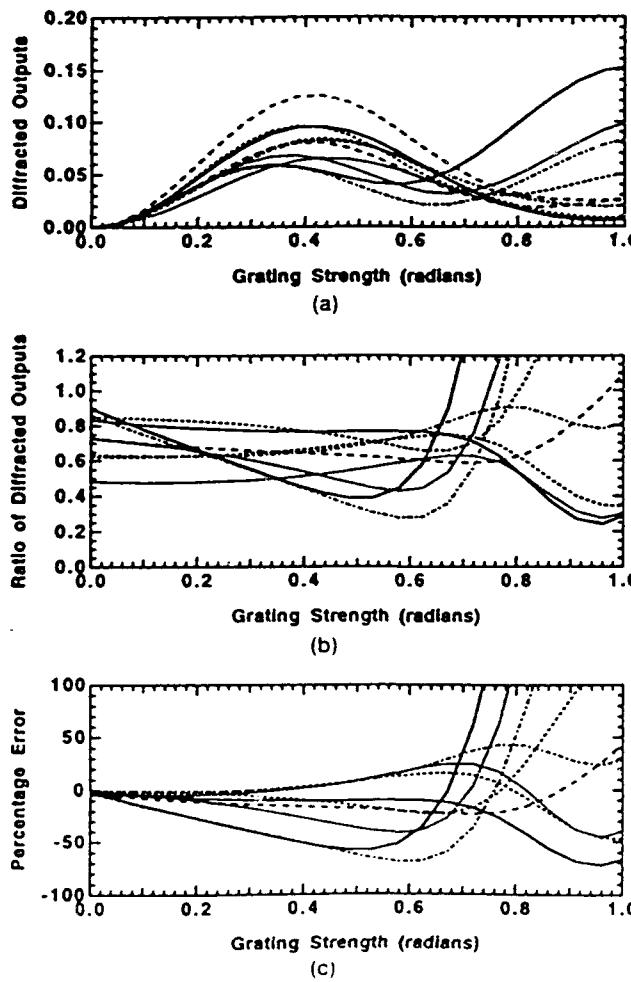


Fig. 7. Same as Fig. 5 but for the pagewise-sequential recording method with a beam splitter ratio  $R$  of 1000.

Because increasing the beam splitter ratio can reduce the effects of the cross gratings for the pagewise-sequential recording method, one might ask whether increasing the beam splitter ratio for the simultaneous-recording method in the single-source architecture offers any benefits. As can be seen by comparing Eqs. (24)–(26) for the grating strengths of the interconnection cross gratings for the simultaneous-recording method, increasing the beam splitter ratio reduces the relative magnitudes of the training-plane cross gratings by a factor of  $\sqrt{R}$  [Eq. (26)] as compared with the desired interconnection gratings [Eq. (24)]. However, the relative magnitudes of the input-plane cross gratings increase by a factor of  $\sqrt{R}$  [Eq. (25)]. The net result is that varying the beam splitter ratio from unity always enhances one set of cross gratings relative to the desired set of interconnection gratings. Our simulations show that this results in significantly reduced fidelity and throughput performance. Thus a unity beam splitter ratio is optimal for the simultaneous recording method.

### C. Fully Sequential Recording and Readout with Mutually Coherent Beams

Fully sequential recording with  $M$  training pairs results in a refractive-index distribution in the holographic medium of

$$\Delta n(\mathbf{r}) = \sum_{i=1}^N \sum_{j=1}^N \Delta n_{ij}^{(M)} \cos(\mathbf{K}_{ij} \cdot \mathbf{r}), \quad (34)$$

in which there are no undesired cross gratings. The corresponding grating strengths are given by Eq. (24).

Simulation results for readout with mutually coherent beams are shown in Fig. 8. The beam splitter ratio  $R$  is assumed to be unity during recording since there are no cross gratings to minimize. Each diffracted output [Fig. 8(a)] appears more sinusoidal in nature as a function of grating strength than that for either the pagewise-sequential or simultaneous recording cases discussed above. However, the first peak (and the following minimum) of each diffracted output occurs at a different grating strength, with the result that the reconstruction fidelity is somewhat grating-strength dependent [as indicated in Fig. 8(b)]. This result is similar to the behavior noted in Fig. 5 of Ref. 19 for a 5-to-5 interconnection, which was modeled numerically using a coupled-wave approach. The percentage error of the ratios, shown in Fig. 8(c), shows a significant improvement over the pagewise-sequential and simultaneous recording cases that is attributable to elimination of the cross gratings. At the peak throughput of >85% (see Fig. 6) the largest ratio percentage error is 15%. For 10% throughput the largest error is only 3%.

Although the fidelity and throughput performance are greatly improved when the cross gratings are eliminated, the fully sequential recording method has several serious difficulties for practical systems, including large numbers of recording steps, inefficient

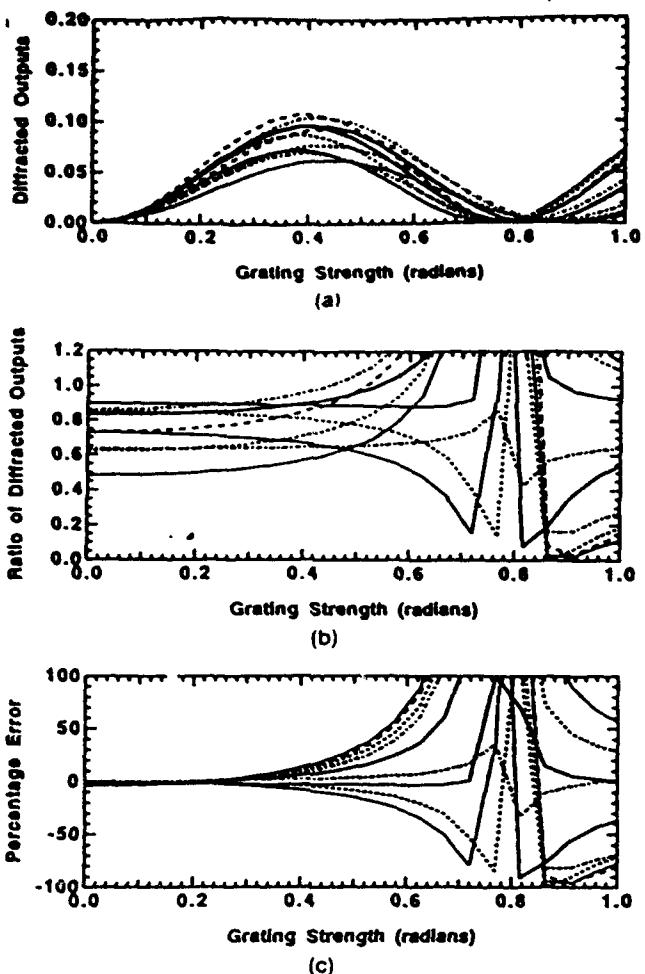


Fig. 8. Single-source architecture simulation results for the pagewise-sequential recording method for a beam splitter ratio,  $R$ , of 1000: (a) diffracted outputs, (b) ratio of each diffracted output to an arbitrarily chosen output, and (c) percentage error of each ratio. Readout is performed with mutually coherent beams.

use of the illuminating beams, and complicated recording schedules (see Section 2.D).

#### D. Readout with Mutually Incoherent Beams

To this point in the discussion we have considered readout of single-source architectures with mutually coherent beams only. In order to better understand the innovations incorporated in the incoherent/coherent double angularly multiplexed architecture discussed in Section 5, we briefly discuss readout of the single-source architecture with mutually *incoherent* beams. For the purposes of the section we consider the fully sequential recording method, such that only the desired interconnection gratings are present in the holographic medium.

Simulation results for the diffracted outputs in this case are shown in Fig. 9. The reconstruction fidelity is essentially the same as for readout with mutually coherent beams (Fig. 8), even though the ideal outputs are expressed in this case by Eq. (4) instead of Eq. (6). Since beam degeneracy is always present in the single-source architecture, the low fidelity error

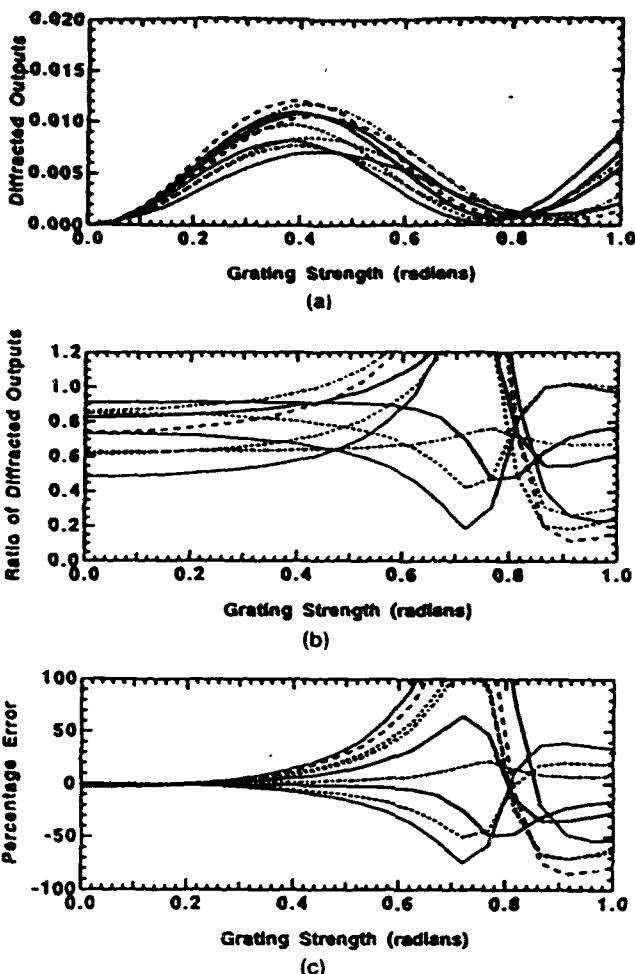


Fig. 9. Single-source architecture simulation results for the fully sequential recording method: (a) diffracted outputs, (b) ratio of each diffracted output to an arbitrarily chosen output, and (c) percentage error of each ratio. Readout is performed with mutually coherent beams.

results shown for both mutually incoherent and coherent readout beams indicate that beam degeneracy is not a significant source of fidelity error (at least, not for grating strengths up to  $\sim 0.4$  rad, at which the peak in throughput occurs). However, for readout with mutually incoherent beams the peak throughput drops to 10% (as shown in Fig. 6) because of incoherent fan-in loss, which limits the maximum throughput to  $1/N$  for an  $N$ -to- $N$  interconnection system with collinear output summation.<sup>24</sup>

An understanding of the physical mechanism responsible for incoherent fan-in loss in a single-source holographic interconnection system permits us to design a system that avoids this loss (as discussed in Section 5). The fan-in loss in a single-source architecture for readout with incoherent beams can be explained by the *beam degeneracy* that is present in the diffracted outputs.<sup>9,10,13,25</sup> Beam degeneracy refers to the  $k$ -vector degeneracies in beams diffracted from different gratings in holographic interconnection systems that have collinear fan-in. This can be understood by reference to Fig. 10(a), in which a

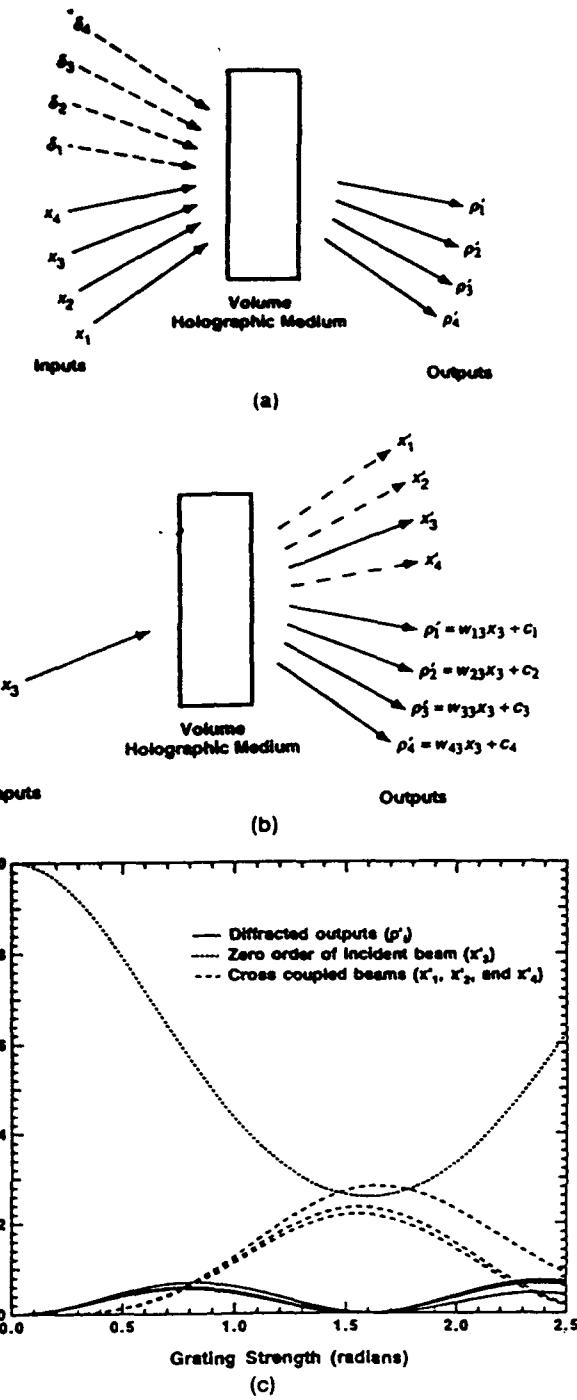


Fig. 10. (a) Schematic diagram of the recording geometry of a 4-to-4 interconnection system; (b) schematic diagram of the readout geometry of the 4-to-4 interconnection system with a single beam ( $x_3$ ) and the resulting outputs; (c) simulation results for readout of the 4-to-4 interconnection system, in which the power cross coupled by beam degeneracy from the desired outputs to the  $x_1'$ ,  $x_2'$ , and  $x_4'$  beams can be significant.

4-to-4 interconnection system is represented schematically. As in Fig. 4, each collimated beam is represented by its  $k$  vector. When readout is performed using all four input beams ( $x_1$ - $x_4$ ), four reconstructed beams emerge collinearly in each of the four  $p_i'$  directions. Thus the four beams that are diffracted

in a given  $p_i'$  direction have degenerate wave vectors; i.e., they are *degenerate beams*.

To understand how this degeneracy leads to a fan-in loss for readout with mutually incoherent beams, consider readout of the 4-to-4 interconnection system described above with a single input beam,  $x_3$ . As shown by the solid arrows in Fig. 10(b), diffracted outputs are generated in each of the four output directions ( $p_1$ '- $p_4$ ') as well as the zero-order direction  $x_3'$ . However, each diffracted beam  $p_i'$  is in turn diffracted into the directions  $x_1'$ ,  $x_2'$ , and  $x_4'$  by the gratings recorded among the training beams  $\delta_1$ - $\delta_4$  and the other input beams  $x_1$ ,  $x_2$ , and  $x_4$ . These interactions cause power to be coupled out of the desired outputs  $p_i'$  and into the zero orders of  $x_1$ ,  $x_2$ , and  $x_4$ , generating the cross-coupled beams  $x_1'$ ,  $x_2'$ , and  $x_4'$ .

The magnitude of this effect can be illustrated by modeling a 4-to-4 interconnection system (using the BPM) in which a single beam of unit intensity [corresponding to  $x_3$  in Fig. 10(b)] is used for readout. Uniform weights are assumed ( $W_{ij} = 1$  for all  $i$  and  $j$ ), and the geometry of the system used in the modeling is the same as that described in Section 3.C.1. The results of the BPM calculation are shown in Fig. 10(c), in which the diffraction efficiencies of all of the diffracted beams are shown as a function of grating strength. The diffraction efficiencies of the desired outputs  $p_i'$  are indicated with solid curves, while the zero-order and cross-coupled beams are represented by the dotted and dashed curves, respectively. The peak diffraction efficiency into the desired outputs  $p_i'$  occurs at  $\sim 0.8$  rad. The summed power in the four desired outputs is only 24%, which is  $1/4$  (or  $1/N$ ) of the maximum available power. The rest of the power either remains in the zero order ( $x_3'$ ) or is diffracted into the cross-coupled beams ( $x_1'$ ,  $x_2'$ , and  $x_4'$ ), each of which has approximately the same diffraction efficiency as the desired beams (at 0.8 rad).

It is clear that if the other three input beams ( $x_1$ ,  $x_2$ , and  $x_4$ ) are also incident and that if all of the incident beams are mutually incoherent, then the summed diffraction efficiency into the desired outputs is still only 25% at best, because the light in each output beam adds incoherently. The end result in this case is analogous to that corresponding to the collinear combination of  $N$  mutually incoherent quasi-monochromatic beams of essentially identical wavelengths using  $N - 1$  beam splitters; only  $1/N$  of the power at most can be diverted into the desired direction, while the remaining power exits the system in the direction of the original input beams.<sup>13</sup>

However, as illustrated in the simulation results of Section 4.C, it is possible to achieve high optical throughput in the single-source architecture if the incident beams are mutually coherent and the proper phase conditions are satisfied. Alternatively, if a holographic interconnection system is designed such that each detected output consists of *angularly* instead of *collinearly* fanned-in beams, the throughput loss that is due to the effects of beam degeneracy can

be avoided even for readout with mutually incoherent beams. The incoherent/coherent double angularly multiplexed architecture discussed in Section 5 illustrates one method of achieving such angular fan-in at each pixel in the output plane with consequent high throughput for readout with mutually incoherent beams.

## E. Discussion

### 1. Comparison of Recording Methods Using an rms Fidelity Error

The percentage-error ratio metric used above to assess the reconstruction fidelity of the single-source architecture under various recording conditions is attractive in that it graphically illustrates the variation in fidelity among the individual diffracted outputs. However, a lumped error measure is more convenient for comparing not only different recording methods for the single-source architecture but also for comparing the reconstruction fidelity of different architectures.

One possible error measure is given by<sup>19</sup>

$$\epsilon = |\hat{u} - \hat{u}'| = \left[ \sum_{j=1}^N \left( \frac{\rho_j}{|\rho|} - \frac{\rho'_j}{|\rho'|} \right)^2 \right]^{1/2}, \quad (35)$$

in which  $\hat{u} = \rho/|\rho|$  is the unit vector in the direction of  $\rho$ ;  $\hat{u}'$  is defined similarly. The error measure  $\epsilon$  can be interpreted as the rms error of the components of the normalized diffracted output vector  $\hat{u}'$ . The separate normalizations of the ideal and actual components by  $|\rho|$  and  $|\rho'|$ , respectively, permit a change in the throughput that does not in turn bias the relative fidelity of the components. The maximum value of the rms error  $\epsilon$  is  $\sqrt{2}$  since  $\hat{u}$  and  $\hat{u}'$  are unit vectors having unipolar components.

The rms errors of the various recording methods discussed in Sections 4.A–4.D for the 10-to-10 simulations of the single-source architecture are shown in Fig. 11 as a function of the grating strength of the largest interconnection grating. For simultaneous recording the rms error starts out relatively small at essentially zero grating strength (and practically no throughput, as shown in Fig. 6), and then becomes large quite rapidly as the grating strength is increased. If just the cross gratings among the input-plane pixels are removed using pagewise-sequential recording with  $R = N^2 = 100$  (i.e., only the desired interconnection gratings and the training-plane cross gratings are present, with comparable grating strengths), neither the fidelity nor the throughput improves significantly. Decreasing the magnitudes of the training-plane cross gratings relative to the desired interconnection gratings (using pagewise-sequential recording with  $R = 1000$ ) results in substantial improvement in both the fidelity and the peak throughput.

Complete elimination of the cross gratings by using sequential recording yields a further significant improvement in fidelity and a marginal increase in the peak throughput, as shown in Fig. 6. The reconstruc-

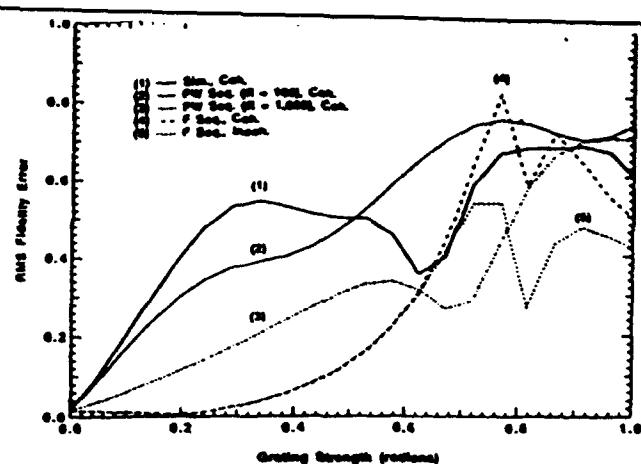


Fig. 11. Simulation results showing the rms fidelity error [as defined in Eq. (35)] for several combinations of recording and readout methods for the single-source architecture. An explanation of the legend is provided in Fig. 6.

tion fidelity that we obtain in the absence of cross gratings is much higher than that shown by Slinger.<sup>19</sup> This difference can be understood as follows. Slinger's analysis considers both random complex weights (i.e., each interconnection grating has a random phase as well as a random amplitude) and random-amplitude readout beams that have either a 0° or 180° relative phase. The direct implication is that the readout beams do not have the same relative phases as the recording beams that were used to create the weights [i.e., the argument of the exponential in Eq. (7) is not constant for all  $i$  and  $j$ , such that condition (1) discussed following Eq. (6) in Section 2.C is violated]. The net result is poor reconstruction fidelity.

In our analysis of the single-source interconnection architecture the readout beams have the same relative phases as the recording beams, which permits the appropriate phasing conditions to be satisfied upon reconstruction such that good fidelity results. This same-relative-phase condition is consistent with the implementation of unipolar weights and outputs, as is discussed in Section 2.C (recall that a dual-rail strategy can be generalized to the bipolar case instead of requiring the interconnection system to implement bipolarity directly). A comparison of our results with Slinger's indicates that (in the absence of cross gratings) a single-source architecture can in fact yield high reconstruction fidelity for an appropriate mapping of neural interconnection requirements to the architecture.

### 2. Effects of Sidelobe Overlap on Reconstruction Fidelity

A curious feature of the simulation results shown in Fig. 11 is that the fidelity error for each recording method does not go to zero with decreasing grating strength. Furthermore, the asymptotic value of the fidelity error near zero grating strength is the same for each recording method. This indicates that the

limiting value of the error is independent of the presence or absence of cross gratings.

The cause of this somewhat counterintuitive behavior appears to be diffraction from non-Bragg-matched interconnection gratings.<sup>48</sup> The physical mechanism for this effect is illustrated in Fig. 12, in which the +1-order diffraction efficiency of two sinusoidal phase gratings is shown as a function of the incidence angle of a plane-wave readout beam. The gratings are assumed to have the same grating period, but their fringes are oriented at different slant angles relative to the front face of the holographic medium. If a readout beam is incident at  $\theta_1$  (Bragg matched to grating  $G_1$ ), it is diffracted by both  $G_1$  (through  $G_1$ 's mainlobe response) and grating  $G_2$  (through  $G_2$ 's sidelobe response at  $\theta_1$ ). Since  $G_1$  and  $G_2$  have the same grating period, the light beams diffracted from both gratings are collinear upon exiting the holographic medium. Hence the overall diffraction response at incidence angle  $\theta_1$  is composed of a small contribution from  $G_2$  as well as the main contribution from  $G_1$ . Upon coherent addition of the two contributions the net diffraction efficiency in general differs from what it would be in the absence of  $G_2$ . If  $G_1$  implements a weighted interconnection, this difference results in a small error in the weight of the grating. We refer to the source of this error as sidelobe overlap.

In the single-source architecture illustrated in Figs. 3 and 4, there are multiple sets of gratings that have the same grating period but different slant angles because of the existence of multiple pairs of writing beams originating from the pixels of the input and training planes that have equal angular separations. For example, the grating written by beams  $x_1$  and  $\delta_1$  (Fig. 4) has the same grating period (but a different slant angle) as the grating written by beams  $x_2$  and  $\delta_2$ . As a result of the presence of sidelobe overlap among

various sets of gratings, one would expect the weights of the affected gratings to exhibit a small residual error upon reconstruction. This variation in the recorded weights should be present even as the grating strength goes to zero because both the main Bragg response of a particular grating and the overlapping sidelobes of the other gratings scale in proportion to each other as a function of grating strength.

One method of testing the assertion that sidelobe overlap affects the reconstruction fidelity as described above is to vary the thickness of the holographic medium such that the degree of sidelobe overlap among the various interconnection gratings is changed (since the angular width of the diffraction response of each grating is inversely proportional to the thickness<sup>49</sup>). If sidelobe overlap is present, one would expect the limiting value of the rms fidelity error to generally increase with decreasing thickness, and vice versa. This behavior is exactly what we observe in our simulations.

For the simultaneous and pagewise-sequential cases shown in Fig. 11, the effect of sidelobe overlap is apparent only at small grating strengths because the cross gratings are the main source of fidelity error at larger grating strengths. For the fully sequential recording case with coherent readout, sidelobe overlap appears to be the dominant source of reconstruction error up to  $\sim 0.25$  rad. As borne out by other simulation results, the reconstruction fidelity in this regime can be further improved simply by increasing the thickness of the holographic medium.

### 3. Scaling Trends

To examine how our modeling results scale with the number of nodes in the interconnection system, we also simulated a 4-to-4 single-source interconnection architecture having a  $4 \times 4$  weight matrix that is a subset of the  $10 \times 10$  weight matrix used above. The corresponding four input beams from  $\mathbf{x}^{\text{read}}$  are used to read out the interconnection system. Simulation results for the rms fidelity error of the normalized output vectors and the throughput for the 4-to-4 case are shown in Figs. 13(a) and 13(b), respectively, for the simultaneous, pagewise-sequential, and fully sequential recording methods.

The overall performance characteristics noticed in the 10-to-10 simulations are also present in the 4-to-4 results, in that the rms fidelity errors and throughputs for the various cases show no identifiable trends for scale-up from the 4-to-4 case to the 10-to-10 case. As a specific example, the simulation results for the relative behaviors of the pagewise-sequential and simultaneous recording methods in the 4-to-4 case are similar to the 10-to-10 results (see Fig. 13). The rms fidelity error for the simultaneous recording method in both cases increases rapidly as the grating strength increases from zero. When the input-plane cross gratings are removed (i.e., when pagewise-sequential recording is used with  $R = 16$  for the 4-to-4 case and  $R = 100$  for the 10-to-10 case), neither the fidelity nor the throughput improves significantly

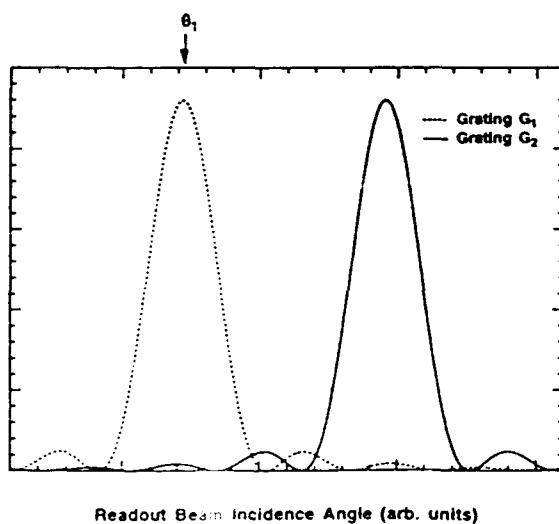


Fig. 12. Angular response characteristics of two Bragg gratings that have the same grating period and slightly different slant angles. Although the main lobes of the angular responses are well separated, the sidelobes and the main lobes overlap.

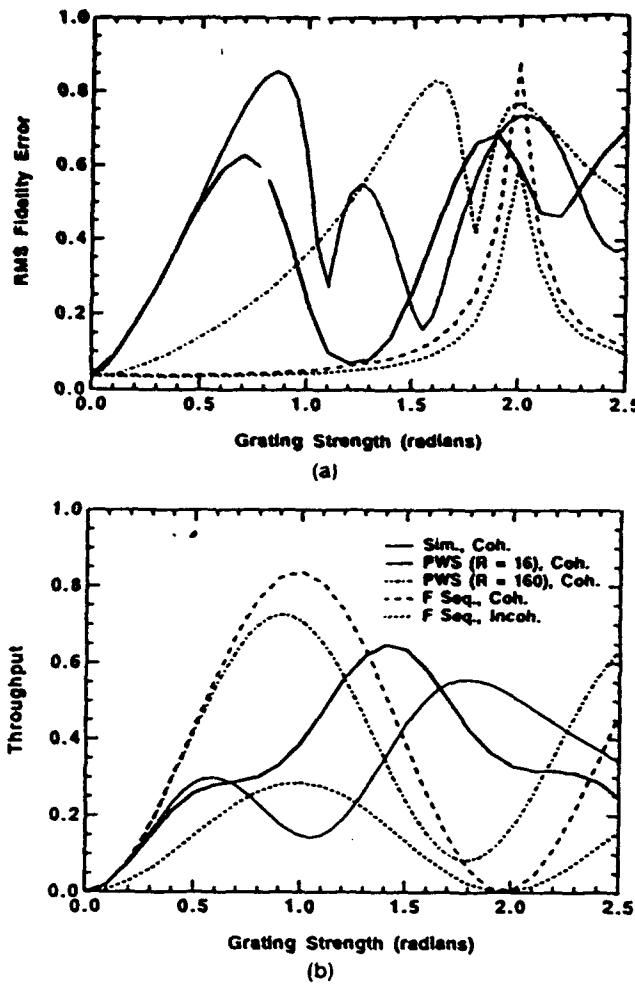


Fig. 13. Simulation results for a 4-to-4 single-source architecture. Shown are (a) the rms fidelity error for various recording and readout combinations and (b) the optical throughput for the various recording and readout combinations. Clarification of the legend [in (b), from top to bottom] is as follows: simultaneous recording method with coherent readout, pagewise-sequential recording method with a beam splitter ratio of 16 and coherent readout, pagewise-sequential recording method with a beam splitter ratio of 160 and coherent readout, fully sequential recording with coherent readout, and fully sequential recording with incoherent readout.

in either case. Increasing the strength of the desired interconnections to approximately three times the training-plane cross gratings (i.e., when pagewise-sequential recording is used with  $R = 160$  for the 4-to-4 case) reduces the fidelity error and significantly increases the peak throughput, just as in the 10-to-10 case. In fact, the rms fidelity error is nearly the same for both simulations at a grating strength corresponding to the peak throughput in each case. At the current level of simulation complexity it is not yet clear whether this apparent insensitivity to the dimensions of the interconnection system is generalizable.

In addition to these similarities, there are a few interesting differences among these particular 4-to-4 and 10-to-10 cases. For simultaneous recording the

4-to-4 case shows a significant drop in the rms fidelity error at  $\sim 1.2$  rad (which happens to correspond to 50% throughput). This result is weight-matrix and input-vector dependent and cannot be relied on to occur in general. Similar comments apply to the observed drop in fidelity error for the  $R = 16$  pagewise-sequential recording case that occurs at grating strengths in excess of 0.85 rad. Despite these differences, the general trends do not appear to be weight-matrix and input-vector dependent.

Furthermore, as the grating strength goes to zero the fidelity error for the 4-to-4 case [Fig. 13(a)] is the same for all of the recording methods, as in the 10-to-10 cases. However, the actual value of the error is somewhat smaller for the 10-to-10 results than for the 4-to-4 simulations. This difference is probably caused by the increased number of overlapping sidelobes that affect a particular weighted interconnection in the 10-to-10 case. As the number of overlapping sidelobes increases, their effect on the interconnection fidelity may tend to decrease because the sign of each sidelobe's contribution to the weighted interconnection can be either positive or negative, depending on which particular sidelobe is accessed for the off-Bragg grating. The contributions of a large number of sidelobes may therefore tend to average to zero. The implication is that the limiting fidelity error that is due to sidelobe overlap should decrease as the number of nodes in the interconnection architecture increases (at least until some other limiting phenomenon is reached). This should result in better fidelity performance for the sequential recording method for throughputs at which the fidelity is limited by sidelobe overlap.

#### F. Single-Source Architecture Simulation Results: Conclusions

Our simulation results show that, as expected, the coherent-recording cross-talk gratings for simultaneous recording in a single-source architecture cause a significant degradation in reconstruction fidelity for reasonable throughputs. Pagewise-sequential recording shows better fidelity and peak-throughput performance than simultaneous recording if the beam splitter ratio can be made large enough to overcome the undesired strengthening of the training-plane cross gratings, which is caused by the larger number of exposures that they receive relative to the desired interconnection gratings. Sequential recording yields both high reconstruction fidelity and high optical throughput, but at the cost of  $N^2 - 1$  more recording steps than for the case of simultaneous recording.

Based on the considerations discussed in this paper, an attractive method of implementing a single-source architecture is to employ simultaneous recording in a geometry that clearly separates the range of spatial frequencies obtained for the desired interconnection gratings from those obtained for the cross gratings; and to use a holographic material that is sensitive to the former range of spatial frequencies and insensitive to the latter.<sup>17,19</sup> Of course this

assumes that high throughput is required in the resultant interconnection system. No special requirements are placed on the spatial-frequency sensitivity of the holographic medium if low throughputs are permissible. In this regime the simultaneous recording method yields the same fidelity performance as the other recording methods, all of which are limited by the amount of angular sidelobe overlap that is present.

Even if a single-source architecture is implemented as described above, it has at least one remaining characteristic that detracts from its potential use in implementations of large-scale weighted fan-out/fan-in interconnection systems. As mentioned in Section 2.E, the density of pixels on the input and training planes is limited to a certain degree by grating degeneracy.<sup>33,34</sup>

### 5. Incoherent/Coherent Double Angularly Multiplexed Interconnection Architecture

In this section we discuss the operation of the incoherent/coherent double angularly multiplexed architecture that we have proposed and investigated recently<sup>8-14</sup> and compare its relative merits to those of single-source architectures. Two configurations of the incoherent/coherent double angularly multiplexed architecture are presented: first, the full-aperture configuration<sup>8-14</sup> and second, the subhologram configuration.<sup>11,12,14,50</sup> Both configurations permit simultaneous recording of each training pair with significantly reduced coherent-recording cross talk as compared with the single-source architecture. In addition, readout is performed with mutually incoherent beams such that, during operation, each diffracted output is described by the usual neural-network summation of Eq. (4) rather than the modified summation of Eq. (6). This is accomplished without sacrificing the high throughput efficiency that is typically associated only with fully coherent systems, because both configurations avoid the presence of beam degeneracy. Furthermore, while grating degeneracy is present in the full-aperture configuration of the architecture, it can potentially be eliminated by using the subhologram configuration and therefore does not require subsampling of the input and training planes.

The full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture is described in Section 5.A, and simulation results are discussed in Section 5.B. Operation of the subhologram configuration of the architecture is discussed in Section 5.C, and simulation results follow in Section 5.D. Section 5.E provides a comparison of the two configurations.

#### A. Full-Aperture Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Operation

A schematic diagram for one layer of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture is shown in Fig. 14(a). The architecture has three key components:

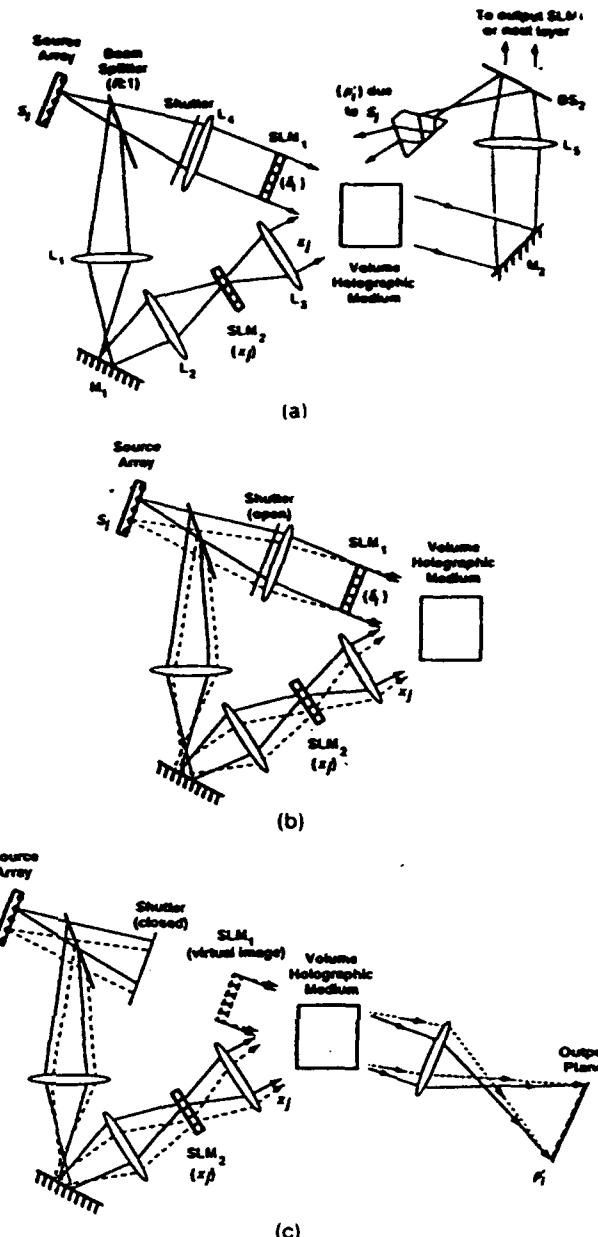


Fig. 14. Schematic diagram of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture showing (a) general layout, (b) recording, and (c) reconstruction.  $M_1$  and  $M_2$  are mirrors;  $L_1$ - $L_3$  are lenses;  $BS_2$  is a second beam splitter.

(1) a 2-D array of individually coherent but mutually incoherent sources; (2) optoelectronic neuron-unit arrays that integrate the functions of light detection, neuron-unit nonlinear response, and optical modulation for each pixel (these arrays are denoted as  $SLM_1$  and  $SLM_2$ ); and (3) the volume holographic interconnection medium. In this section, we briefly review the interconnection method used in the architecture. Additional details are provided in Refs. 12 and 13.

The process of recording a set of interconnections is illustrated in Fig. 14(b). Assume for the moment that a single pair of training vectors  $[x^{(m)}, \delta^{(m)}]$  is to be recorded. Light from each source  $S_j$  is split into two

optical paths. In the upper path each beam is collimated and illuminates the full aperture of the training plane ( $SLM_1$ ). Each beam therefore reads out the training vector  $\delta^{(m)}$  that is present on  $SLM_1$  (each of these beams is referred to herein as a  $\delta^{(m)}$  beam). The collimated beams deriving from the full set of sources propagate at different angles through the training plane. In the lower path the source array is imaged onto the input plane ( $SLM_2$ ) such that light from each source illuminates only one corresponding pixel. A collimating lens converts the beams emerging from the pixels of  $SLM_2$  to a set of angularly distinct collimated beams that illuminate the full aperture of the volume holographic medium.

As a result of the individually coherent but mutually incoherent nature of the sources in the source array, interconnection gratings are formed only between each  $x_j^{(m)}$  and the corresponding  $\delta^{(m)}$  beam with which it is mutually coherent. This permits the simultaneous recording of a set of angularly multiplexed holograms in which each hologram is formed by the interference of an angularly distinct reference beam  $x_j^{(m)}$  with a second angularly distinct beam (from source  $S_j$ ) bearing the image  $\delta^{(m)}$ . The full set of image-bearing beams  $[\{\delta^{(m)}\}]_j$  that derives from all of the sources in the source array  $[S_j]$  and that encodes the contents of  $SLM_1$  is also angularly multiplexed. Hence we describe this architecture as *double angularly multiplexed*.<sup>13</sup>

Similar to the case of simultaneous recording in a single-source architecture, the incoherent/coherent double angularly multiplexed architecture requires only one exposure to record each training pair  $[\mathbf{x}^{(m)}, \delta^{(m)}]$ , which is accomplished by turning on all of the sources in the source array simultaneously. Since mutually incoherent beams are used to read out the pixels of  $SLM_2$ , no coherent-recording cross-talk gratings among the input-plane pixels are formed. Similarly, cross gratings among the separate beams encoded with  $\delta^{(m)}$  do not occur. The only cross gratings that can form in the holographic medium result from overlaps among adjacent diffracted components  $\delta_i$  within each  $\delta^{(m)}$  beam in the Fresnel regime. Depending on the size of the pixels and the distance between  $SLM_1$  and the holographic medium, these cross gratings connect any single pixel only to those in some local neighborhood of the pixel. The effects of such local cross talk can be minimized by adjusting the beam splitter ratio  $R$ . This interconnection system therefore permits simultaneous recording of each training pair while minimizing the effects of coherent-recording cross talk.

As illustrated in Fig. 14(c), readout is performed using the lower optical path (with all of the sources turned on simultaneously). The volume holographic optical element, or VHOE, performs the requisite set of weighted fan-outs, while the imaging lens following the VHOE performs an optical fan-in operation by imaging the diffracted beams onto the pixels of the output plane. The angularly distinct set of collimated beams that illuminated  $SLM_1$  during recording

is therefore reconstructed by the VHOE; after passing through the lens, the beams form a real image in the output plane, which is conjugate to the  $SLM_1$  plane. The net result is that a fan-in of angularly distinct incoherent beams is performed at each node in the output plane. As long as the angular spread is sufficiently large, an incoherent fan-in can be performed without incurring the usual fan-in loss associated with a collinear incoherent fan-in. If we use appropriate optical elements (depending on the particular neural-network model being implemented), the output plane shown in Fig. 14(c) may be coincident with the input side of  $SLM_1$  itself [as shown in Fig. 14(a)], with  $SLM_2$ , with the input  $SLM$  of the next layer, or with any combination thereof.

#### B. Full-Aperture Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Simulation

The optical beam propagation method was used to analyze the fidelity and throughput performance of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture by simulating a 10-to-10 interconnection system in which the same weight matrix was recorded as above for the single-source architecture and in which the same input vector  $\mathbf{x}^{read}$  was used for readout. In all such simulations, readout was performed using mutually incoherent beams.

To facilitate direct comparison with the single-source architecture simulations, we assume that the same holographic medium characteristics (linear material with a thickness of 4.5 mm and a refractive index of 2.52) and the same operating wavelength (0.514  $\mu m$ ) are used. The separation of the sources in the source array is the same as the pixel separation in  $SLM_1$  and  $SLM_2$ , which is 257  $\mu m$ . Instead of using two lenses to image the source array onto  $SLM_2$  [lenses  $L_1$  and  $L_2$  in Fig. 14(a)], we use a single lens only, which is separated by twice its focal length from both the source array and  $SLM_2$ . The focal lengths of lenses  $L_4$  and  $L_3$  are assumed to be 50 mm. The separation between  $SLM_1$  and the holographic medium is also assumed to be 50 mm. The beam splitter ratio  $R$  is unity.

Given the above parameters, light passing through each pixel of  $SLM_1$  spreads approximately 200  $\mu m$  laterally in propagating to the holographic medium because of diffraction. Since this is less than one pixel width, we assume for simplicity that geometrical optics adequately describes the propagation of light from  $SLM_1$  to the holographic medium. Our model thus does not consider the effects of any diffraction-induced local-neighborhood cross gratings that may be present. Instead, each beam that reads out  $SLM_1$  [referred to as a  $\delta^{(m)}$  beam above] is assumed to propagate essentially unchanged to the holographic medium, forming an exact (rather than an approximate) image of  $SLM_1$  on its front face. The interference of each  $\delta^{(m)}$  beam with light from its associated  $x_j^{(m)}$  pixel in  $SLM_2$ , therefore results in ten (for a 10-to-10 interconnection) distinct grating regions in

the holographic medium. In each distinct region the resulting interconnection grating pattern has a different amplitude, proportional to  $|x_i^{(M)} \delta_i^{(M)}|^2$  for a single training pair, while the grating periods and slant angles are the same for all ten regions that derive from source  $S_i$ .

For the parameters assumed above for the source and pixel separations, for the focal length of  $L_4$ , and for the separation between SLM<sub>1</sub> and the holographic medium, each  $\delta_i^{(M)}$  beam is shifted by one pixel from angularly adjacent  $\delta_i^{(M)}$  beams on the face of the hologram. The net result is 19 separate regions in the holographic medium, with between 1 and 10 gratings multiplexed in each region. The 2-D formulation of the BPM discussed in Section 3.3 was used to simulate readout of each distinct region.

Simulation results for the diffracted outputs of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture are shown in Fig. 15 as a function of the grating strength of the largest interconnection grating. The rms fidelity error and throughput are shown in Fig. 16; despite readout by a set of mutually incoherent beams, the peak throughput is nearly 95%.

As in our fidelity analysis of the single-source architecture, the functional dependence of each weight on its corresponding grating strength determines the ideal input/output proportionality factor that must be known for comparison with the actual input/output characteristics of the system. For holographic interconnection systems the ideal input/output relationship is based on the underlying physics of the diffraction process used in the system.

For example, direct application of the weight relationship expressed in Rel. (29) for a single-source architecture to the fidelity analysis of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture suggests the presence of significant fidelity errors, as illustrated in Figs. 15(c) and 16(a). Although the percentage errors in Fig. 15(c) are small for low throughput, they become quite large (up to 85%) at the peak throughput. Similar behavior is observed for the rms fidelity error in Fig. 16(a) (solid curve).

The reason for the apparent lack of fidelity is that Rel. (29) does not adequately describe the physics behind the diffraction process used in the full-aperture configuration. In this case each grating in a given region of the holographic material is completely independent of the other interconnection gratings in that region, except for effects such as angular sidelobe overlap. We therefore assume that the diffraction efficiency of any particular grating is given by  $\sin^2(\nu_{ij}/2)$ , in which  $\nu_{ij}$  is the strength of this grating.<sup>49</sup> On this basis the corresponding weight relationship for the full-aperture configuration of the architecture (for multiple training pairs) is

$$W_{ij} \propto \sin^2[\nu_{ij}^{(M)}/2], \quad (36)$$

in which  $\nu_{ij}^{(M)}$  is given by Eq. (24).

Comparison of the diffracted outputs obtained from

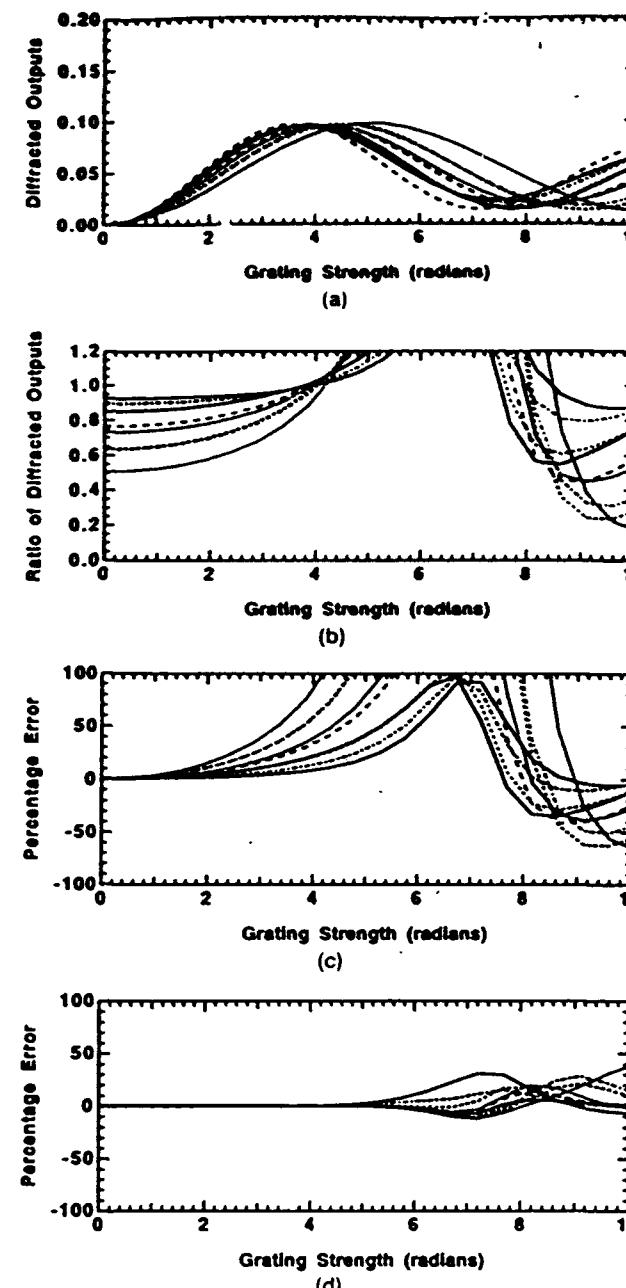


Fig. 15. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (full-aperture configuration) for readout with mutually incoherent beams. Shown as functions of the grating strength of the largest grating are (a) the diffracted outputs, (b) the ratios of the diffracted outputs, (c) the percentage error using Rel. (28) for the dependence of each weight on grating strength, and (d) the percentage error using Rel. (36) for the dependence of each weight on grating strength.

the BPM simulations to the ideal outputs calculated using Rel. (36) [as shown in Figs. 15(d) and 16(a)] yields much better measured fidelity performance. The actual diffracted outputs from the holographic medium are of course not changed; rather, the metric against which they are compared is related more closely to the underlying diffraction behavior of the interconnection system. The weight definition of

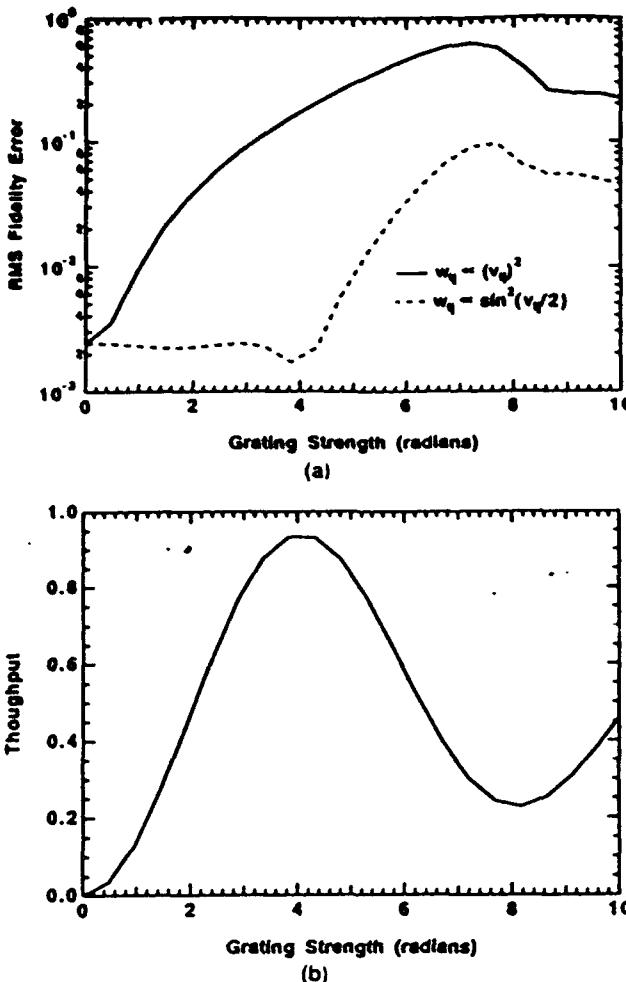


Fig. 16. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (full-aperture configuration). Shown are (a) the rms fidelity error for two different functional dependencies of the weights on the grating strength and (b) the optical throughput.

Rel. (36) as compared with that of Rel. (29) implies a different functional form for the weight-update relationship in an adaptive system. The effect of this altered functional form on the performance of learning algorithms is currently unknown. We conjecture that, given a physical implementation with limited dynamic range for each weight, the soft-limiting characteristic provided by the  $\sin^2$  function may prove to be in some respects preferable to a hard-clipping saturation characteristic.

As shown in Fig. 15(d), the percentage errors of the ratios of the diffracted outputs obtained using the  $\sin^2$  weight relationship are very close to zero for throughputs up to and including the peak throughput of 95%. This result shows that the principal source of error in this configuration indeed derives from the metric rather than from some source of cross talk. The comparison of the two metrics is illustrated clearly in Fig. 16(a), in which the rms fidelity errors are plotted on a log scale. The rms fidelity error for the  $\sin^2$  metric is relatively flat between 0 and 4 rad of

grating strength (at which the peak throughput occurs) and does not go to zero as  $v \rightarrow 0$ . The principal reason for the nonzero fidelity error throughout this region appears to be overlap of the angular sidelobes of the gratings multiplexed in each region of the hologram. Increasing the number of interconnection nodes in the incoherent/coherent double angularly multiplexed architecture should lower this level of error by averaging out the contributions of an increased number of sidelobes, as discussed in Section 4.E.3.

For the 10-to-10 interconnection system analyzed, the rms fidelity error shown in Fig. 16(a) with a weight relationship given by Rel. (36) is dramatically smaller than for the sequentially recorded single-source architecture using either mutually coherent or incoherent readout beams. For example, at the grating strengths corresponding to peak throughput, the error for the incoherent/coherent double angularly multiplexed architecture is more than an order of magnitude lower than that for the sequentially recorded case of the single-source architecture.

For completeness we applied the weight relationship of Rel. (36) to the fidelity analysis of the single-source architecture and found that the fidelity errors were essentially the same as shown in Fig. 11. This result is not surprising, because in the single-source architecture the grating strength of each individual grating is small (0.4 rad for the largest interconnection grating at the peak throughput), and Rel. (36) reduces to Rel. (29) for small  $v_{ij}^{(M)}$ .

In certain volume holographic media, such as photorefractive crystals, the presence of many overlapping incoherent beams in the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture results in a small modulation depth for each pair of recording beams, which in turn significantly reduces the achievable optical throughput.<sup>12</sup> This problem may be avoided by using the subhologram configuration of the incoherent-coherent double angularly multiplexed architecture, which is discussed in Section 5.C.

Furthermore, as mentioned in the beginning of Section 5, the full-aperture configuration of the incoherent-coherent double angularly multiplexed architecture is additionally subject to the effects of grating degeneracy. By modifying the architecture slightly to realize the subhologram configuration, we can potentially avoid this source of cross talk without using fractal sampling grids. The trade-off, however, is permitting the presence of additional cross gratings. As discussed below, these can in turn be minimized by adjusting the beam splitter ratio.

### C. Subhologram Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Operation

As shown in Fig. 17, the subhologram configuration of the double angularly multiplexed architecture can be created by inserting an additional lens ( $L_5$ ) between  $SLM_1$  and the holographic medium. The lens

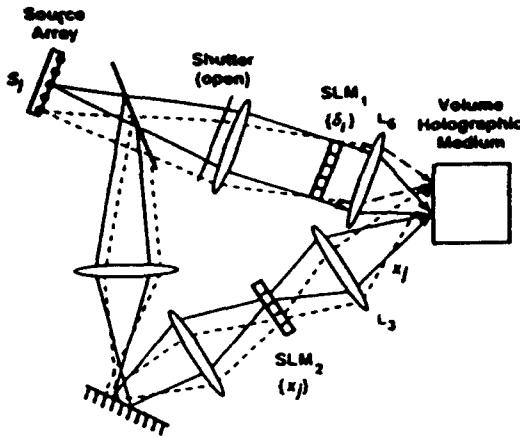


Fig. 17. Schematic diagram of the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture.

is positioned at a distance of one focal length from the medium such that it performs a Fourier transform of the beams emerging from SLM<sub>1</sub> (this corresponds to imaging the source array onto the medium). In addition, lens L<sub>3</sub> is adjusted to image SLM<sub>2</sub> onto the medium, which also effectively reimages the source array instead of collimating the light from each pixel in SLM<sub>2</sub> as in the full-aperture configuration of the architecture.

The optical system is designed such that the images of the source array through the upper and lower paths of the architecture are in registry at the holographic medium so that spatially distinct holograms (i.e., subholograms) are formed across the face of the medium (as shown schematically in Fig. 18). To understand the nature of each subhologram, let us focus on only one source S<sub>j</sub> that is imaged onto the hologram through both paths. The image of S<sub>j</sub> through the upper path contains the Fourier transform of the image of  $\delta^{(m)}$  that is on SLM<sub>1</sub>, while the image of S<sub>j</sub> through the lower path has an intensity proportional to  $x_j^{(m)}$ . The interference between the two beams creates weighted interconnections between the jth pixel in the input plane and all of the training-plane pixels. However, interference among

the components of the Fourier transform of  $\delta^{(m)}$  causes cross gratings that form intraplanar connections among the training-plane pixels. By adjusting the beam splitter ratio R, we can decrease the magnitude of these cross gratings relative to the desired interconnection gratings.

Each of the subholograms connects a single pixel in the input plane to all of the pixels in the training plane and thus performs a 1-to-N weighted fan-out upon reconstruction. As in the full-aperture configuration of the architecture, an imaging lens is used after the holographic medium [shown as L<sub>5</sub> in Fig. 14(a)] to perform the fan-in to each node in the output plane. The subholograms in general will at least partially overlap within the volume holographic medium, depending on the focal lengths of the lenses and the spacings of the pixels and of the sources. However, this spatial overlap does not cause additional cross gratings to form during simultaneous exposure of the set of subholograms because the sources are mutually incoherent. During recording, all of the sources are turned on simultaneously such that the recording of M training pairs requires only M exposures, just as in the full-aperture configuration of the architecture. Full illumination of both SLM apertures is accomplished with the entire source array on, which provides efficient power transfer to the holographic medium during each exposure.

In Section 5.D we discuss simulation results and scaling trends for the subhologram configuration of the double angularly multiplexed architecture.

#### D. Subhologram Configuration of the Incoherent/Coherent Double Angularity Multiplexed Architecture: Simulation

In order to determine the relationship between an individual weight and the strength of its associated grating within the subhologram configuration of the incoherent-coherent double angularly multiplexed architecture, we use the fact that each spatially segregated subhologram implements an independent 1-to-N fan-out. As discussed in Section 3.B, an analytical solution for the diffraction efficiency of a 1-to-N weighted fan-out has been obtained using coupled-wave theory under the assumption that no cross gratings are present.<sup>38</sup> The net result is that the weights and grating strengths for the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture are related by (in our notation)

$$W_{ij} \propto [v_{ij}^{(M)}]^2, \quad (37)$$

which is the same as Rel. (29) for the single-source architecture. The fidelity analysis presented in this section is based on the use of Rel. (37) to compute the ideal input/output characteristics of the subhologram configuration.

For the simulations discussed in this section, the weight matrix and the readout vector used are the same as those described in Section 3.C.3. Also, the parameters of the optical components are chosen to

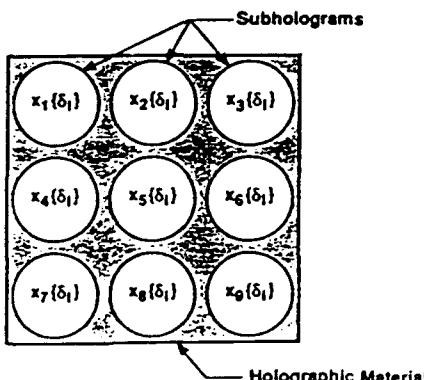


Fig. 18. Schematic diagram of a subhologram array. Each subhologram is shown as spatially separate in this case.

be the same as for the full-aperture configuration of the architecture, except that the focal length of  $L_4$  is 25 mm, that of  $L_3$  is 100 mm, and that of  $L_2$  is 50 mm. For the purposes of our modeling, the resultant subholograms ( $\sim 500 \mu\text{m}$  in diameter, set on  $500-\mu\text{m}$  centers) are considered to be fully separated, and the pixels in the training plane are treated as point sources. The optical beam propagation method was used to model the readout of each subhologram.

Simulation results are shown in Figs. 19 and 20. In Fig. 19 the individual diffracted outputs are computed for a beam splitter ratio  $R$  of 100. The horizontal axis, as above, refers to the grating strength of the largest interconnection grating in the holographic medium. The ratio percentage errors in Fig. 19(c) are not much different than those shown in Fig. 8(c) for sequential recording in the single-source architecture. For example, at the peak throughput of over 95% [as shown in Fig. 20(b)] the largest ratio error for the subhologram configuration of the incoherent/coherent double angularly multiplexed

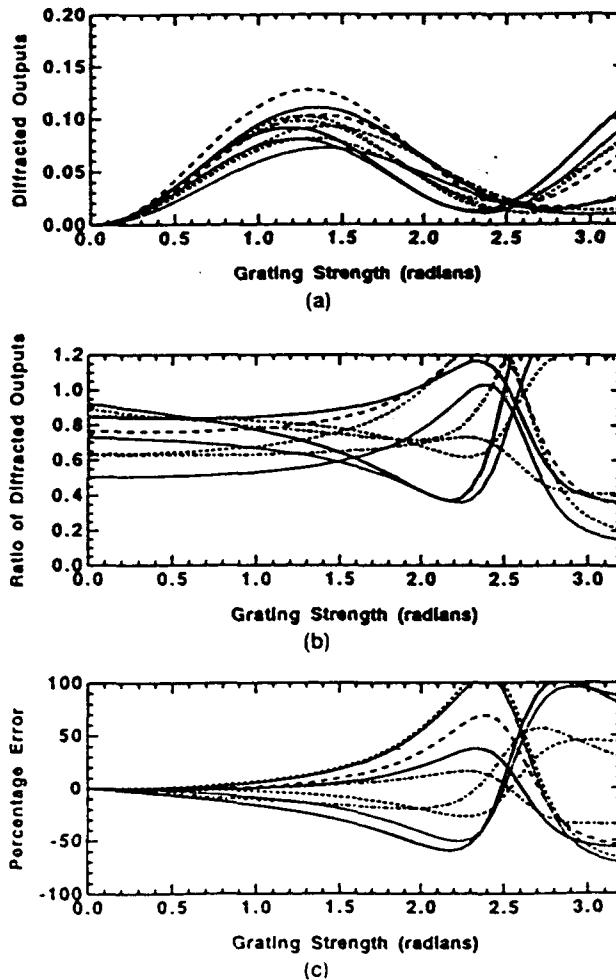


Fig. 19. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (subhologram configuration) for readout with mutually incoherent beams. Shown as functions of the grating strength of the largest grating are (a) the diffracted outputs, (b) the ratios of the diffracted outputs, and (c) the percentage error of each ratio.

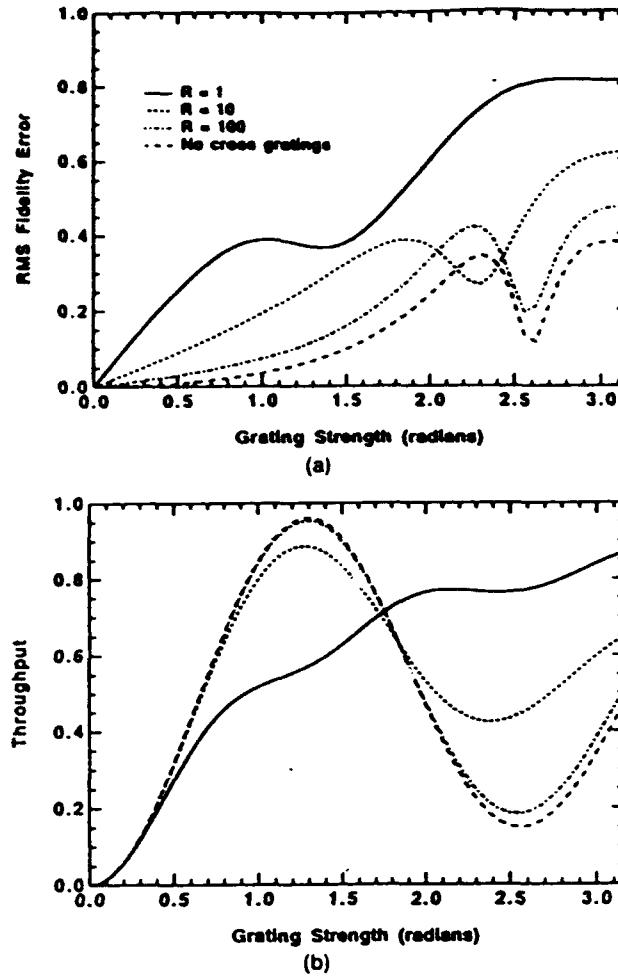


Fig. 20. (a) The rms fidelity error and (b) the throughput for various beam splitter ratios in the 10-to-10 incoherent-coherent double angularly multiplexed architecture (subhologram configuration). When  $R = 100$ , the fidelity error and throughput approach the case for which there are no cross gratings. Readout is performed with mutually incoherent beams.

architecture (with  $R = 100$ ) is 25%; at 10% throughput, the largest ratio error is only 4%.

The rms fidelity error for the subhologram configuration is shown in Fig. 20(a) for several values of the beam splitter ratio, and the corresponding throughputs are shown in Fig. 20(b). When the beam splitter ratio is unity (i.e., when the cross gratings have the same relative amplitudes as the desired interconnection gratings), the cross gratings cause large fidelity errors except in the limit of low throughput (similar to the results obtained for the single-source architecture). As the relative magnitudes of the cross gratings decrease with increasing beam splitter ratios, the fidelity and throughput both improve. For comparison, a case is also shown for which no cross gratings are present. In all cases, the fidelity error asymptotically approaches zero with decreasing grating strength because angular sidelobe overlap that might affect the desired interconnections does not occur in the subhologram configuration.

In order to examine how these results scale with the number of interconnection nodes in the training input planes, we simulated a 4-to-4 interconnection system using the same weight matrix and readout vector as in the single-source architecture 4-to-4 simulations. The fidelity error and throughput results are shown in Fig. 21 for the same beam splitter ratios as shown in Fig. 20. For  $R = 100$  neither the fidelity nor the throughput seems to differ significantly between the 10-to-10 and the 4-to-4 results. For the case  $R = 1$ , however, substantial improvement in both fidelity and throughput is observed in scaling up to the 10-to-10 interconnection system from the 4-to-4 system (except at small grating strengths).

#### Discussion of the Incoherent/Coherent Double Angularly Multiplexed Architecture Configurations

Our simulations demonstrate the ability of the coherent/coherent double angularly multiplexed architecture to obtain high optical throughput (at least for linear holographic materials) when mutually

incoherent beams are used during readout of the holographic interconnections. By avoiding beam degeneracy in both configurations of the architecture, we can circumvent the usual incoherent fan-in loss found in the single-source architecture.

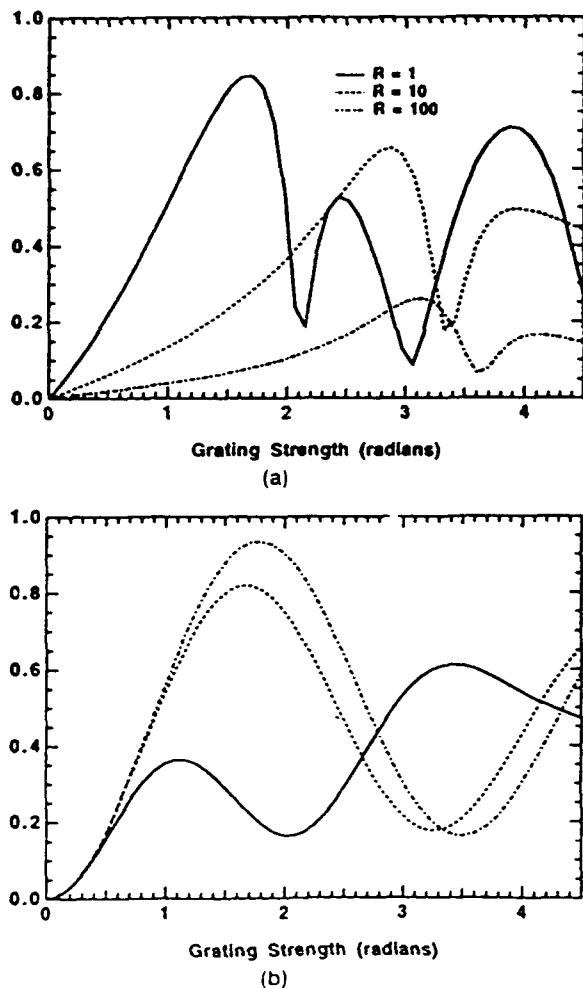
Our simulation results (presented in Section 5.B) further demonstrate that high reconstruction fidelity is achievable in the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture for a fidelity metric based on the diffraction properties of its interconnection system. As noted above, grating degeneracy is present in the full-aperture configuration such that fractal sampling grids may be required in certain applications. By contrast, the subhologram configuration of the architecture avoids the presence of grating degeneracy, which may permit an increased interconnection density for a given physical system volume relative to both the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture and the single-source architecture. However, the subhologram configuration involves a fundamental trade-off between reconstruction fidelity and the beam splitter ratio resulting from the presence of cross gratings that are not present in the full-aperture configuration.

An important aspect of the subhologram configuration is the incorporation of both spatial and angular multiplexing in the holographic medium to obtain independence of the interconnection gratings. In the limiting case of complete spatial separation of the subholograms, only spatial multiplexing is used. In this case, a thin holographic material could in principle be used in the interconnection system. However, since large numbers of interconnections ( $10^8$ - $10^{10}$ ) are anticipated for photonic neural networks, space-bandwidth limitations will in general necessitate some degree of subhologram overlap in order for compact system implementations to be realized. In this case, the independence of the interconnection gratings with nonnegligible subhologram overlap necessitates angular (or wavelength) multiplexing to achieve Bragg isolation, which in turn requires the use of a thick holographic medium. The set of overlapping subhologram configurations spans the continuum between the full-aperture and full-subhologram configurations and as such may yield an optimum compromise between these two extremes. In fact, the optimal degree of subhologram overlap may well prove to be material dependent.

#### 6. Comparison of Holographic Interconnection Techniques

In this section we compare and contrast the full-aperture and subhologram configurations of the incoherent/coherent double angularly multiplexed architecture with the single-source architecture configured using different recording methods.

The fidelity and throughput performance of the single-source architecture and of both configurations



21 (a) The rms fidelity error and (b) the throughput for various beam splitter ratios in the 4-to-4 incoherent/coherent double angularly multiplexed architecture (subhologram configuration). Comparison with Fig. 20 indicates how fidelity and throughput variations scale with the number of interconnection nodes.

of the incoherent/coherent double angularly multiplexed architecture are summarized in Fig. 22, in which the rms fidelity errors for the 10-to-10 simulations are shown as a function of optical throughput instead of grating strength. For each curve in the figure the right-hand end point represents the peak throughput achieved in the simulation for the particular interconnection architecture and recording method to which that curve corresponds.

As a result of the effects of cross gratings, use of the simultaneous recording method in the single-source architecture yields both poor reconstruction fidelity and a peak throughput of 50% (for the case simulated). For a beam splitter ratio of 100 (the same as shown for the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture) the pagewise-sequential recording method does not yield significant performance improvement. As mentioned in Section 4.B, a serious drawback for the use of this recording method is that the beam splitter ratio required for a given level of fidelity error increases quadratically with the number of nodes in the interconnection system. When all of the cross gratings are eliminated in the single-source architecture by using a sequential recording technique, both high throughput and good reconstruction fidelity are achievable.

As the pagewise-sequential and fully sequential recording methods within the single-source architecture require a significantly larger number of exposure steps per training pair and greater hardware complexity than the simultaneous recording method, they are potentially less attractive options for implementation of large-scale adaptive neural-network systems.

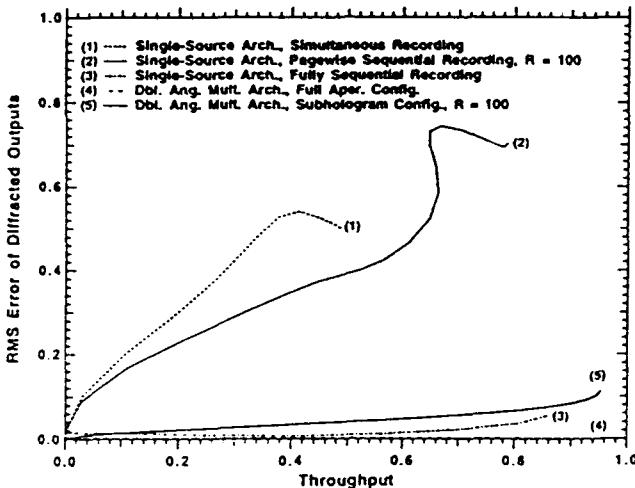


Fig. 22. Rms error as a function of the throughput for the single-source interconnection architecture (parameterized by recording method) and for the two configurations of the incoherent/coherent double angularly multiplexed architecture. In all cases, the single-source architecture is read out with mutually coherent beams and the incoherent/coherent double angularly multiplexed architecture is read out with mutually incoherent beams. The curve for the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture lies almost directly on the horizontal axis.

Alternatively, simultaneous recording in a single-source architecture suffers from a lack of reconstruction fidelity for significant optical throughput. If low throughput is tolerable in a given computational architecture, or if the effects of the cross gratings can be minimized by using the spatial-frequency-sensitive properties of a particular holographic material, a single-source architecture with mutually coherent readout beams and an appropriate fractal sampling grid becomes a viable option.

As illustrated in Fig. 22, both configurations of the incoherent/coherent double angularly multiplexed architecture can achieve both high fidelity and high optical throughput while using simultaneous recording (with a large enough beam splitter ratio in the case of the subhologram configuration). High optical throughput proves to be obtainable (in a linear holographic medium) despite the use of mutually incoherent readout beams because each configuration avoids the presence of beam degeneracy.

In addition to providing for linear summation of the diffracted output intensities, readout with mutually incoherent beams in the incoherent/coherent double angularly multiplexed architecture avoids (during operation, not training) the rigid optical phase stability requirements needed in a single-source architecture that is read out with mutually coherent beams. This feature reduces the degree of vibration isolation required and hence increases the practicality of operating a trained photonic neural network in an industrial or field environment.

A further advantage of the incoherent/coherent double angularly multiplexed architecture is that the interconnection gratings in the volume holographic medium can be copied into a second volume holographic recording medium in a single recording step.<sup>51</sup> For example, the full set of interconnections that are learned in a primary adaptive system can easily be reproduced in any number of secondary permanent holographic media for operational use. In contrast, direct single-step copying of an interconnection pattern within the single-source architecture is not possible without sacrificing either interconnection fidelity or optical throughput. Instead, it appears that at least  $N$  (if not  $N^2$ ) exposure steps are required for duplication of an  $N$ -to- $N$  interconnection system within a single-source architecture.

As mentioned in Section 5.E, greater interconnection densities may be achievable if grating degeneracy (and hence the use of fractal sampling grids) can be avoided. Of the interconnection techniques discussed herein, only the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture offers the potential of avoiding the presence of grating degeneracy.

In this paper we have quantitatively evaluated the performance characteristics of the incoherent/coherent double angularly multiplexed architecture (based on the use of an array of individually coherent but mutually incoherent sources) for highly multiplexed volume holographic interconnection ap-

plications. In addition, we have quantitatively evaluated the directly comparable performance characteristics of conventional single-source architectures. As discussed below, there are several clear directions for continuing research.

### 7. Future Research Directions

In this study we have performed a detailed comparison of a number of holographic interconnection architectures that can be used to implement weighted interconnections with a high degree of fan-out and fan-in. For the most part, this comparison has been made on the basis of interconnection pathway independence (lack of cross talk) and insertion loss (optical throughput efficiency). For the neural-network application in particular, it would be of considerable interest to determine the appropriate levels of inter-channel isolation and insertion loss permissible in the context of particular learning models without compromising overall system performance. In other words, to what degree are certain neural-network models sensitive (or insensitive) to these effects, if we rely to a greater or lesser extent on the learning capacity of the network to obviate the necessity for ideal interconnection behavior? Preliminary experimental and theoretical studies suggest the ability of some learning algorithms to overcome a certain degree of cross talk in the interconnection system,<sup>52,53</sup> but a more comprehensive study of this issue is necessary.

The simulation studies presented herein should be expanded to evaluate the additional limitations imposed by the effects of self-diffraction among the recording beams, grating erasure, exposure scheduling, finite pixel size, and finite range of grating strength modulation on both reconstruction fidelity and throughput, particularly as the number of interconnections is increased. Significantly increasing the number of interconnection nodes considered in this analysis will enable scaling trends for the relative errors of each architecture to be further identified and compared. Inclusion of the grating recording characteristics of photorefractive media in the holographic-recording model will permit the effects of material nonlinearities to be determined and the utility of these materials for the implementation of adaptive photonic neural networks to be evaluated. Furthermore, extension of the BPM simulations to three dimensions will permit verification of the trends observed using a two-dimensional model and will also permit direct investigation of grating-degeneracy effects.

In addition to further modeling studies, previous laboratory work that has confirmed the basic features of the incoherent/coherent double angularly multiplexed architecture<sup>12</sup> can be expanded to include a more detailed study of various implementation issues, particularly as applied to photorefractive media. Such issues include the quantitative comparison of experimentally determined fidelity errors and optical throughput losses with simulation results; the effects of scale-up in the number of nodes on interconnection

performance; the effects of subhologram overlap and the beam splitter ratio on reconstruction fidelity and throughput; and continued device development that will permit the eventual integration of mutually compatible source arrays, neuron-unit arrays, and volume holographic media into a practical system.

This research was supported in part by the Defense Advanced Research Projects Agency (both the Program on Neural Networks and the National Center for Integrated Photonic Technology) and the U.S. Air Force Office of Scientific Research (University Research Initiative: Center for the Integration of Optical Computing). Greg Nordin gratefully acknowledges support from a Hughes Doctoral Fellowship and a Center for Photonic Technology Graduate Prize Fellowship at the University of Southern California.

### References

1. R. Kostuk, J. Goodman, and L. Hesselink, "Design considerations for holographic optical interconnects," *Appl. Opt.* **26**, 3947-3953 (1987).
2. D. Z. Anderson and D. M. Lininger, "Dynamic optical interconnects: volume holograms as optical two-port operators," *Appl. Opt.* **26**, 5031-5038 (1987).
3. D. Psaltis, D. J. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," *Appl. Opt.* **27**, 1752-1758 (1988).
4. M. Cronin-Golomb, "Dynamically programmable self-aligning optical interconnect with fan-out and fan-in using self-pumped phase conjugation," *Appl. Phys. Lett.* **54**, 2189-2191 (1989).
5. J. H. Hong, S. Campbell, and P. Yeh, "Optical pattern classifier with perceptron learning," *Appl. Opt.* **29**, 3019-3025 (1990).
6. E. S. Maniloff and K. M. Johnson, "Dynamic holographic interconnects using static holograms," *Opt. Eng.* **29**, 225-229 (1990).
7. See, for example, the feature on neural networks, *Appl. Opt.* **26**, 4909-5111 (1987).
8. B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay, Jr., "Photonic architecture for neural nets using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest* (ICO-90 Organizing Committee, Kobe, Japan, 1990).
9. P. Asthana, H. Chin, G. P. Nordin, A. R. Tanguay, Jr., S. Piazzolla, and B. K. Jenkins, "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest* (ICO-90 Organizing Committee, Kobe, Japan, 1990).
10. P. Asthana, H. Chin, G. P. Nordin, A. R. Tanguay, Jr., G. C. Petrisor, B. K. Jenkins, and A. Madhukar, "Photonic components for neural net implementations using incoherent-coherent holographic interconnections," in *1990 OSA Annual Meeting*, Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 57.
11. B. K. Jenkins, A. R. Tanguay, Jr., S. Piazzolla, G. C. Petrisor, and P. Asthana, "Photonic neural-network architecture based on incoherent-coherent holographic interconnections," in *1990 OSA Annual Meeting*, Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 56.
12. P. Asthana, "Volume holographic techniques for highly multiplexed interconnection applications," Ph.D. dissertation (University of Southern California, Los Angeles, Calif., 1991).
13. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic implementa-

-tion. of neural networks," in *Neural Networks for Signal Processing*, B. K. Laskin, ed. (Prentice-Hall, Englewood Cliffs, N.J., 1992), Chap. 9, pp. 287-382.

14. G. P. Nordin, "Volume diffraction phenomena for photonic neural network implementations and stratified volume holographic optical elements," Ph.D. dissertation (University of Southern California, Los Angeles, Calif., 1992).
15. D. Psaltis, A. A. Yamamura, K. Hsu, S. Lin, X.-G. Gu, and G. Brady, "Optoelectronic implementations of neural networks," *IEEE Commun. Mag.* **27**(11), 37-40 (1989).
16. H. Lee, "Volume holographic global and local interconnecting patterns with maximal capacity and minimal first-order cross talk," *Appl. Opt.* **28**, 5312-5316 (1989).
17. H. Lee, X.-G. Gu, and D. Psaltis, "Volume holographic interconnections with maximal capacity and minimal cross talk," *J. Appl. Phys.* **65**, 2191-2194 (1989).
18. C. X.-G. Gu, "Optical neural networks using volume holograms," Ph.D. dissertation (California Institute of Technology, Pasadena, Calif., 1990).
19. C. Slinger, "Analysis of the  $N$ -to- $N$  volume-holographic neural interconnect," *J. Opt. Soc. Am. A* **8**, 1074-1081 (1991).
20. J. A. Fleck, J. R. Morris, and M. D. Feit, "Time-dependent propagation of high energy laser beams through the atmosphere," *Appl. Phys.* **10**, 129-160 (1976).
21. J. Van Roey, J. van der Donk, and P. E. Lagasse, "Beam-propagation method: analysis and assessment," *J. Opt. Soc. Am.* **71**, 803-810 (1980).
22. D. Yevick and L. Thylen, "Analysis of gratings by the beam-propagation method," *J. Opt. Soc. Am.* **72**, 1081-1089 (1982).
23. R. V. Johnson and A. R. Tanguay, Jr., "Optical beam propagation method for birefringent phase grating diffraction," *Opt. Eng.* **25**, 235-249 (1986).
24. J. W. Goodman, "Fan-in and fan-out with optical interconnections," *Opt. Acta* **32**, 1489-1496 (1985).
25. P. Asthana, G. Nordin, S. Piazzolla, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of interchannel cross talk and throughput efficiency in highly multiplexed fan-out-fan-in holographic interconnections," in *1990 OSA Annual Meeting*, Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 242.
26. D. O. Hebb, *Organization of Behavior* (Wiley, New York, 1949).
27. C. Peterson, S. Redfield, J. D. Keeler, and E. Hartman, "Optoelectronic implementation of multilayer neural networks in a single photorefractive material," *Opt. Eng.* **29**, 359-368 (1990).
28. S. Kessler and R. Hild, "A new method for simultaneous complex addition and subtraction," *Opt. Quantum Electron.* **15**, 65-70 (1983).
29. N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the Hopfield model," *Appl. Opt.* **24**, 1469-1475 (1985).
30. G. C. Petrisor, B. K. Jenkins, H. Chin, and A. R. Tanguay, Jr., "Dual-function adaptive neural networks for photonic implementation," in *1990 OSA Annual Meeting*, Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 56.
31. F. H. Mok, M. C. Tackitt, and H. M. Stoll, "Storage of 500 high-resolution holograms in a LiNbO<sub>3</sub> crystal," *Opt. Lett.* **16**, 605-607 (1991).
32. D. Brady and D. Psaltis, "Control of volume holograms," *J. Opt. Soc. Am. A* **9**, 1167-1182 (1992).
33. D. Psaltis, D. Brady, X.-G. Gu, and K. Hsu, "Optical implementation of neural computers," in *Optical Processing and Computing*, H. Arsenault, ed. (Academic, New York, 1988), pp. 251-276.
34. D. Psaltis, X.-G. Gu, and D. Brady, "Fractal sampling grids for holographic interconnections," in *Optical Computing '88*, P. Chavel, J. W. Goodman, and G. Roblin, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **963**, 468 (1988).
35. E. N. Glytsis and T. K. Gaylord, "Three-dimensional (vector) rigorous coupled-wave analysis of anisotropic grating diffraction," *J. Opt. Soc. Am. A* **7**, 1399-1420 (1990).
36. S. K. Case, "Coupled-wave theory for multiply exposed thick holographic gratings," *J. Opt. Soc. Am.* **65**, 724-729 (1975).
37. L. Solymar, "Two-dimensional  $N$ -coupled-wave theory for volume holograms," *Opt. Commun.* **23**, 199-202 (1977).
38. L. Solymar and D. J. Cooke, *Volume Holography and Volume Gratings* (Academic, New York, 1981), p. 212.
39. C. W. Slinger and L. Solymar, "Volume phase holograms reconstructed by the object wave," *Opt. Quantum Electron.* **16**, 369-372 (1984).
40. E. N. Glytsis and T. K. Gaylord, "Rigorous 3-D coupled wave diffraction analysis of multiple superposed gratings in anisotropic media," *Appl. Opt.* **28**, 2401-2421 (1989).
41. K.-Y. Tu, T. Tamir, and H. Lee, "Multiple-scattering theory of wave diffraction by superposed volume gratings," *J. Opt. Soc. Am. A* **7**, 1421-1435 (1990).
42. K.-Y. Tu, H. Lee, and T. Tamir, "Analysis of cross talk in volume holographic interconnections," *Appl. Opt.* **31**, 1717-1729 (1992).
43. M. D. Feit and J. A. Fleck, Jr., "Light propagation in graded-index optical fibers," *Appl. Opt.* **17**, 3990-3998 (1978).
44. R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography* (Academic, New York, 1971).
45. W. R. Klein and B. D. Cook, "Unified approach to ultrasonic light diffraction," *IEEE Trans. Sonics Ultrason.* **SU-14**, 123-134 (1967).
46. T. K. Gaylord and M. G. Moharam, "Thin and thick gratings: terminology clarification," *Appl. Opt.* **20**, 3271 (1981).
47. B. Benlari and L. Solymar, "The effect of the relative intensity of the reference beam on the reconstructing properties of volume phase gratings," *Opt. Acta* **26**, 271-278 (1979).
48. W. J. Burke and P. Sheng, "Crosstalk noise from multiple thick-phase holograms," *J. Appl. Phys.* **48**, 681-685 (1976).
49. H. Kogelnik, "Coupled wave theory for thick hologram gratings," *Bell Syst. Tech. J.* **48**, 2909-2947 (1969).
50. G. P. Nordin, P. Asthana, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of weighted fan-out/fan-in volume holographic interconnections," in *Diffractive Optics: Design, Fabrication, and Applications*, Vol. 9 of 1992 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1992), pp. 165-167.
51. S. Piazzolla, B. K. Jenkins, and A. R. Tanguay, Jr., "Single-step copying process for multiplexed volume holograms," *Opt. Lett.* **17**, 676-678 (1992).
52. E. G. Paek, J. R. Wullert III, and J. S. Patel, "Holographic implementation of a learning machine based on a multiclass category perceptron algorithm," *Opt. Lett.* **14**, 1303-1305 (1989).
53. C. W. Slinger, "Weighted volume interconnects for adaptive networks," *Opt. Comput. Process.* **1**, 219-232 (1991).